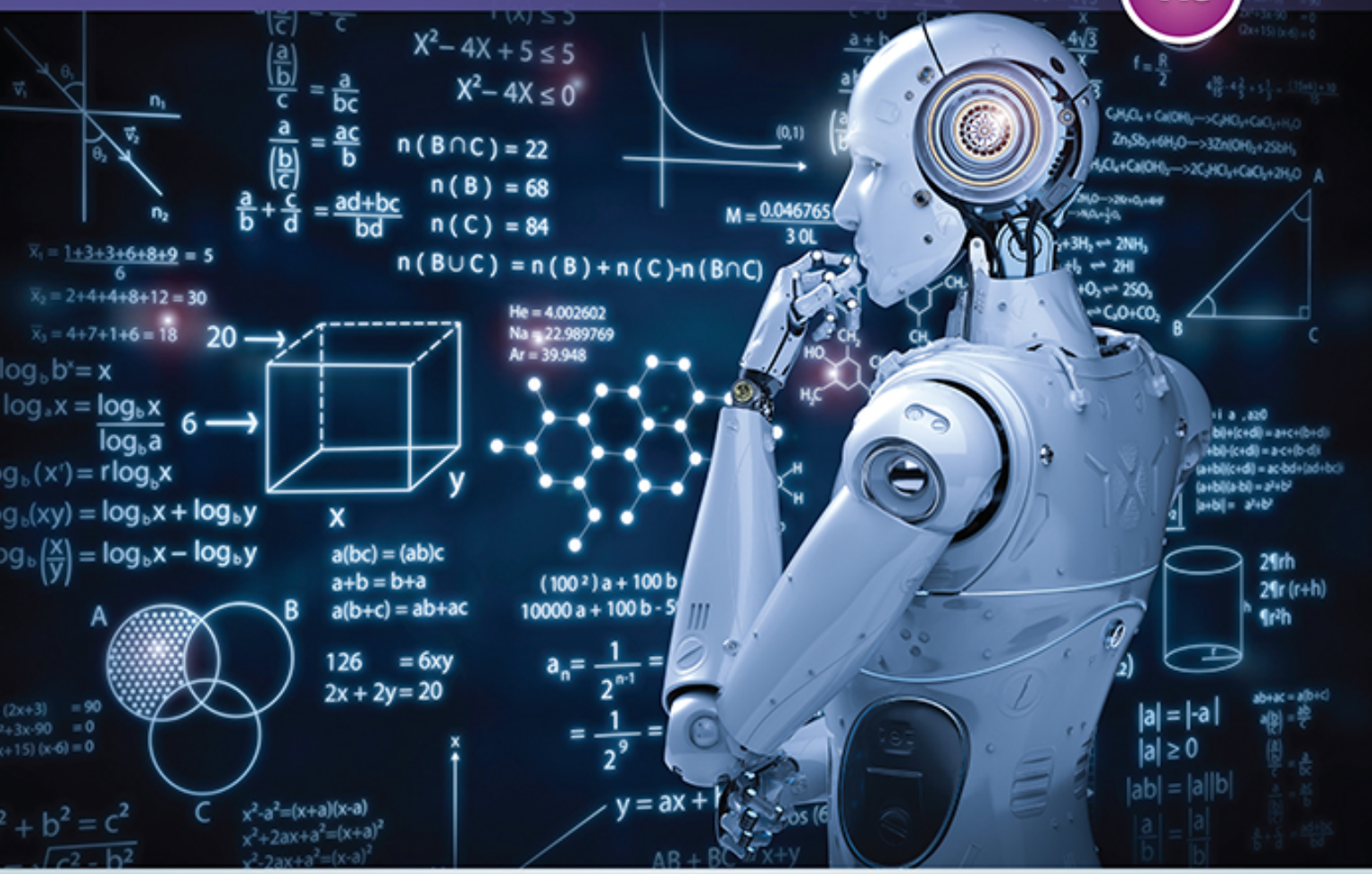
Analytics, Data Science, & Artificial Intelligence

SYSTEMS FOR DECISION SUPPORT

11e



Ramesh Sharda

Dursun Delen

Efraim Turban

ELEVENTH EDITION

ANALYTICS, DATA SCIENCE, & ARTIFICIAL INTELLIGENCE

SYSTEMS FOR DECISION SUPPORT

Ramesh Sharda
Oklahoma State University

Dursun Delen
Oklahoma State University

Efraim Turban
University of Hawaii

Vice President of Courseware Portfolio Management: Andrew Gilfillan
Executive Portfolio Manager: Samantha Lewis
Team Lead, Content Production: Laura Burgess
Content Producer: Faraz Sharique Ali
Portfolio Management Assistant: Bridget Daly
Director of Product Marketing: Brad Parkins
Director of Field Marketing: Jonathan Cottrell
Product Marketing Manager: Heather Taylor
Field Marketing Manager: Bob Nisbet
Product Marketing Assistant: Liz Bennett
Field Marketing Assistant: Derrica Moser
Senior Operations Specialist: Diane Peirano
Senior Art Director: Mary Seiner
Interior and Cover Design: Pearson CSC
Cover Photo: Phonlamai Photo/Shutterstock
Senior Product Model Manager: Eric Hakanson
Manager, Digital Studio: Heather Darby
Course Producer, MyLab MIS: Jaimie Noy
Digital Studio Producer: Tanika Henderson
Full-Service Project Manager: Gowthaman Sadhanandham
Full Service Vendor: Integra Software Service Pvt. Ltd.
Manufacturing Buyer: LSC Communications, Maura Zaldivar-Garcia
Text Printer/Bindery: LSC Communications
Cover Printer: Phoenix Color

Library of Congress Cataloging-in-Publication Data

Library of Congress Cataloging in Publication Control Number: 2018051774

ISBN 10: 0-13-519201-3
ISBN 13: 978-0-13-519201-6

BRIEF CONTENTS

CONTENTS

PREFACE

Analytics has become the technology driver of this decade. Companies such as IBM, Oracle, Microsoft, and others are creating new organizational units focused on analytics that help businesses become more effective and efficient in their operations. Decision makers are using data and computerized tools to make better decisions. Even consumers are using analytics tools directly or indirectly to make decisions on routine activities such as shopping, health care, and entertainment. The field of business analytics (BA)/data science (DS)/decision support systems (DSS)/business intelligence (BI) is evolving rapidly to become more focused on innovative methods and applications to utilize data streams that were not even captured some time back, much less analyzed in any significant way. New applications emerge daily in customer relationship management, banking and finance, health care and medicine, sports and entertainment, manufacturing and supply chain management, utilities and energy, and virtually every industry imaginable.

The theme of this revised edition is analytics, data science, and AI for enterprise decision support. In addition to traditional decision support applications, this edition expands the reader's understanding of the various types of analytics by providing examples, products, services, and exercises by means of introducing AI, machine-learning, robotics, chatbots, IoT, and Web/Internet-related enablers throughout the text. We highlight these technologies as emerging components of modern-day business analytics systems. AI technologies have a major impact on decision making by enabling autonomous decisions and by supporting steps in the process of making decisions. AI and analytics support each other by creating a synergy that assists decision making.

The purpose of this book is to introduce the reader to the technologies that are generally and collectively called *analytics* (or *business analytics*) but have been known by other names such as decision support systems, executive information systems, and business intelligence, among others. We use these terms interchangeably. This book presents the fundamentals of the methods, methodologies, and techniques used to design and develop these systems. In addition, we introduce the essentials of AI both as it relates to analytics as well as a standalone discipline for decision support.

We follow an EEE approach to introducing these topics: **Exposure, Experience**, and **Explore**. The book primarily provides **exposure** to various analytics techniques and their applications. The idea is that a student will be inspired to learn from how other organizations have employed analytics to make decisions or to gain a competitive edge. We believe that such **exposure** to what is being done with analytics and how it can be achieved is the key component of learning about analytics. In describing the techniques, we also introduce specific software tools that can be used for developing such applications. The book is not limited to any one software tool, so the students can **experience** these techniques using any number of available software tools. Specific suggestions are given in each chapter, but the student and the professor are able to use this book with many different software tools. Our book's companion Web site will include specific software guides, but students can gain **experience** with these techniques in many different ways. Finally, we hope that this **exposure** and **experience** enable and motivate readers to **explore** the potential of these techniques in their own domain. To facilitate such **exploration**, we include exercises that direct them to Teradata University Network and other sites as well that include team-oriented exercises where appropriate. In our own teaching experience, projects undertaken in the class facilitate such **exploration** after the students have been **exposed** to the myriad of applications and concepts in the book and they have **experienced** specific software introduced by the professor.

This edition of the book can be used to offer a one-semester overview course on analytics, which covers most or all of the topics/chapters included in the book. It can also be used to teach two consecutive courses. For example, one course could focus on the overall analytics coverage. It could cover selective sections of Chapters 1 and 3–9. A second course could focus on artificial intelligence and emerging technologies as the enablers of modern-day analytics as a subsequent course to the first course. This second course could cover portions of Chapters 1, 2, 6, 9, and 10–14. The book can be used to offer managerial-level exposure to applications and techniques as noted in the previous paragraph, but it also includes sufficient technical details in selected chapters to allow an instructor to focus on some technical methods and hands-on exercises.

Most of the specific improvements made in this eleventh edition concentrate on three areas: reorganization, content update/upgrade (including AI, machine-learning, chatbots, and robotics as enablers of analytics), and a sharper focus. Despite the many changes, we have preserved the comprehensiveness and user friendliness that have made the textbook a market leader in the last several decades. We have also optimized the book's size and content by eliminating older and redundant material and by adding and combining material that is parallel to the current trends and is also demanded by many professors. Finally, we present accurate and updated material that is not available in any other text. We next describe the changes in the eleventh edition.

The book is supported by a Web site (**pearsonhighered.com/sharda**). We provide links to additional learning materials and software tutorials through a special section of the book Web site.

WHAT'S NEW IN THE ELEVENTH EDITION?

With the goal of improving the text and making it current with the evolving technology trends, this edition marks a major reorganization to better reflect on the current focus on analytics and its enabling technologies. The last three editions transformed the book from the traditional DSS to BI and then from BI to BA and fostered a tight linkage with the Teradata University Network (TUN). This edition is enhanced with new materials paralleling the latest trends in analytics including AI, machine learning, deep learning, robotics, IoT, and smart/robo-collaborative assisting systems and applications. The following summarizes the major changes made to this edition.

- ***New organization.*** The book is now organized around two main themes: (1) presentation of motivations, concepts, methods, and methodologies for different types of analytics (focusing heavily on predictive and prescriptive analytic), and (2) introduction and due coverage of new technology trends as the enablers of the modern-day analytics such as AI, machine learning, deep learning, robotics, IoT, smart/robo-collaborative assisting systems, etc. Chapter 1 provides an introduction to the journey of decision support and enabling technologies. It begins with a brief overview of the classical decision making and decision support systems. Then it moves to business intelligence, followed by an introduction to analytics, Big Data, and AI. We follow that with a deeper introduction to artificial intelligence in Chapter 2. Because data is fundamental to any analysis, Chapter 3 introduces data issues as well as descriptive analytics including statistical concepts and visualization. An online chapter covers data warehousing processes and fundamentals for those who like to dig deeper into these issues. The next section covers predictive analytics and machine learning. Chapter 4 provides an introduction to data mining applications and the data mining process. Chapter 5 introduces many of the common data mining techniques: classification, clustering, association mining, and so forth. Chapter 6 includes coverage of deep learning and cognitive computing. Chapter 7 focuses on

text mining applications as well as Web analytics, including social media analytics, sentiment analysis, and other related topics. The following section brings the "data science" angle to a further depth. Chapter 8 covers prescriptive analytics including optimization and simulation. Chapter 9 includes more details of Big Data analytics. It also includes introduction to cloud-based analytics as well as location analytics. The next section covers Robotics, social networks, AI, and the Internet of Things (IoT). Chapter 10 introduces robots in business and consumer applications and also studies the future impact of such devices on society. Chapter 11 focuses on collaboration systems, crowdsourcing, and social networks. Chapter 12 reviews personal assistants, chatbots, and the exciting developments in this space. Chapter 13 studies IoT and its potential in decision support and a smarter society. The ubiquity of wireless and GPS devices and other sensors is resulting in the creation of massive new databases and unique applications. Finally, Chapter 14 concludes with a brief discussion of security, privacy, and societal dimensions of analytics and AI.

We should note that several chapters included in this edition have been available in the following companion book: *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, 4th Edition, Pearson (2018) (Hereafter referred to as BI4e). The structure and contents of these chapters have been updated somewhat before inclusion in this edition of the book, but the changes are more significant in the chapters marked as new. Of course, several of the chapters that came from BI4e were not included in previous editions of this book.

- ***New chapters.*** The following chapters have been added:

  Chapter 2 "Artificial Intelligence: Concepts, Drivers, Major Technologies, and Business Applications" This chapter covers the essentials of AI, outlines its benefits, compares it with humans' intelligence, and describes the content of the field. Example applications in accounting, finance, human resource management, marketing and CRM, and production-operation management illustrate the benefits to business (100% new material)

  Chapter 6, "Deep Learning and Cognitive Computing" This chapter covers the generation of machine learning technique, deep learning as well as the increasingly more popular AI topic, cognitive computing. It is an almost entirely new chapter (90% new material).

  Chapter 10, "Robotics: Industrial and Consumer Applications" This chapter introduces many robotics applications in industry and for consumers and concludes with impacts of such advances on jobs and some legal ramifications (100% new material).

  Chapter 12, "Knowledge Systems: Expert Systems, Recommenders, Chatbots, Virtual Personal Assistants, and Robo Advisors" This new chapter concentrates on different types of knowledge systems. Specifically, we cover new generations of expert systems and recommenders, chatbots, enterprise chatbots, virtual personal assistants, and robo-advisors (95% new).

  Chapter 13, "The Internet of Things as a Platform for Intelligent Applications" This new chapter introduces IoT as an enabler to analytics and AI applications. The following technologies are described in detail: smart homes and appliances, smart cities (including factories), and autonomous vehicles (100% new).

  Chapter 14, "Implementation Issues: From Ethics and Privacy to Organizational and Societal Impacts" This mostly new chapter deals with implementation issues of intelligent systems (including analytics). The major issues covered are protection of privacy, intellectual property, ethics, technical issues (e.g., integration and security) and administrative issues. We also cover the impact of these technologies on organizations and people and specifically deal with the impact on work and

jobs. Special attention is given to possible unintended impacts of analytics and AI (robots). Then we look at relevant technology trends and conclude with an assessment of the future of analytics and AI (85% new).

- ***Streamlined coverage.*** We have optimized the book size and content by adding a lot of new material to cover new and cutting-edge analytics and AI trends and technologies while eliminating most of the older, less-used material. We use a dedicated Web site for the textbook to provide some of the older material as well as updated content and links.
- ***Revised and updated content.*** Several chapters have new opening vignettes that are based on recent stories and events. In addition, application cases throughout the book are new or have been updated to include recent examples of applications of a specific technique/model. These application case stories now include suggested questions for discussion to encourage class discussion as well as further exploration of the specific case and related materials. New Web site links have been added throughout the book. We also deleted many older product links and references. Finally, most chapters have new exercises, Internet assignments, and discussion questions throughout. The specific changes made to each chapter are as follows: Chapters 1, 3–5, and 7–9 borrow material from BI4e to a significant degree.

Chapter 1, "Overview of Business Intelligence, Analytics, Data Science, and Artificial Intelligence: Systems for Decision Support" This chapter includes some material from DSS10e Chapters 1 and 2, but includes several new application cases, entirely new material on AI, and of course, a new plan for the book (about 50% new material).

Chapter 3, "Nature of Data, Statistical Modeling, and Visualization"

- 75% new content.
- Most of the content related to nature of data and statistical analysis is new.
- New opening case.
- Mostly new cases throughout.

Chapter 4, "Data Mining Process, Methods, and Algorithms"

- 25% of the material is new.
- Some of the application cases are new.

Chapter 5, "Machine Learning Techniques for Predictive Analytics"

- 40% of the material is new.
- New machine-learning methods: naïve Bayes, Bayesian networks, and ensemble modeling.
- Most of the cases are new.

Chapter 7, "Text Mining, Sentiment Analysis, and Social Analytics"

- 25% of the material is new.
- Some of the cases are new.

Chapter 8, "Prescriptive Analytics: Optimization and Simulation"

- Several new optimization application exercises are included.
- A new application case is included.
- 20% of the material is new.

Chapter 9, "Big Data, Cloud Computing, and Location Analytics: Concepts and Tools" This material has bene updated substantially in this chapter to include greater coverage of stream analytics. It also updates material from Chapters 7 and 8 from BI4e (50% new material).

Chapter 11, "Group Decision Making, Collaborative Systems, and AI Support" The chapter is completely revised, regrouping group decision support. New topics include

collective and collaborative intelligence, crowdsourcing, swarm AI, and AI support of all related activities (80% new material).

We have retained many of the enhancements made in the last editions and updated the content. These are summarized next:

- ***Links to Teradata University Network (TUN).*** Most chapters include new links to TUN (**teradatauniversitynetwork.com**). We encourage the instructors to register and join **teradatauniversitynetwork.com** and explore the various content available through the site. The cases, white papers, and software exercises available through TUN will keep your class fresh and timely.
- ***Book title.*** As is already evident, the book's title and focus have changed.
- ***Software support.*** The TUN Web site provides software support at no charge. It also provides links to free data mining and other software. In addition, the site provides exercises in the use of such software.

THE SUPPLEMENT PACKAGE: PEARSONHIGHERED.COM/SHARDA

A comprehensive and flexible technology-support package is available to enhance the teaching and learning experience. The following instructor and student supplements are available on the book's Web site, **pearsonhighered.com/sharda**:

- ***Instructor's Manual.*** The Instructor's Manual includes learning objectives for the entire course and for each chapter, answers to the questions and exercises at the end of each chapter, and teaching suggestions (including instructions for projects). The Instructor's Manual is available on the secure faculty section of **pearsonhighered.com/sharda**.
- ***Test Item File and TestGen Software.*** The Test Item File is a comprehensive collection of true/false, multiple-choice, fill-in-the-blank, and essay questions. The questions are rated by difficulty level, and the answers are referenced by book page number. The Test Item File is available in Microsoft Word and in TestGen. Pearson Education's test-generating software is available from **www.pearsonhighered.com/irc**. The software is PC/MAC compatible and preloaded with all of the Test Item File questions. You can manually or randomly view test questions and drag-and-drop to create a test. You can add or modify test-bank questions as needed. Our TestGens are converted for use in BlackBoard, WebCT, Moodle, D2L, and Angel. These conversions can be found on **pearsonhighered.com/sharda**. The TestGen is also available in Respondus and can be found on **www.respondus.com**.
- ***PowerPoint slides.*** PowerPoint slides are available that illuminate and build on key concepts in the text. Faculty can download the PowerPoint slides from **pearsonhighered.com/sharda**.

ACKNOWLEDGMENTS

Many individuals have provided suggestions and criticisms since the publication of the first edition of this book. Dozens of students participated in class testing of various chapters, software, and problems and assisted in collecting material. It is not possible to name everyone who participated in this project, but our thanks go to all of them. Certain individuals made significant contributions, and they deserve special recognition.

First, we appreciate the efforts of those individuals who provided formal reviews of the first through eleventh editions (school affiliations as of the date of review):

Robert Blanning, Vanderbilt University
Ranjit Bose, University of New Mexico

Warren Briggs, Suffolk University
Lee Roy Bronner, Morgan State University
Charles Butler, Colorado State University
Sohail S. Chaudry, University of Wisconsin–La Crosse
Kathy Chudoba, Florida State University
Wingyan Chung, University of Texas
Woo Young Chung, University of Memphis
Paul "Buddy" Clark, South Carolina State University
Pi'Sheng Deng, California State University–Stanislaus
Joyce Elam, Florida International University
Kurt Engemann, Iona College
Gary Farrar, Jacksonville University
George Federman, Santa Clara City College
Jerry Fjermestad, New Jersey Institute of Technology
Joey George, Florida State University
Paul Gray, Claremont Graduate School
Orv Greynholds, Capital College (Laurel, Maryland)
Martin Grossman, Bridgewater State College
Ray Jacobs, Ashland University
Leonard Jessup, Indiana University
Jeffrey Johnson, Utah State University
Jahangir Karimi, University of Colorado Denver
Saul Kassicieh, University of New Mexico
Anand S. Kunnathur, University of Toledo
Shao-ju Lee, California State University at Northridge
Yair Levy, Nova Southeastern University
Hank Lucas, New York University
Jane Mackay, Texas Christian University
George M. Marakas, University of Maryland
Dick Mason, Southern Methodist University
Nick McGaughey, San Jose State University
Ido Millet, Pennsylvania State University–Erie
Benjamin Mittman, Northwestern University
Larry Moore, Virginia Polytechnic Institute and State University
Simitra Mukherjee, Nova Southeastern University
Marianne Murphy, Northeastern University
Peter Mykytyn, Southern Illinois University
Natalie Nazarenko, SUNY College at Fredonia
David Olson, University of Nebraska
Souren Paul, Southern Illinois University
Joshua Pauli, Dakota State University
Roger Alan Pick, University of Missouri–St. Louis
Saeed Piri, University of Oregon
W. "RP" Raghupaphi, California State University–Chico
Loren Rees, Virginia Polytechnic Institute and State University
David Russell, Western New England College
Steve Ruth, George Mason University
Vartan Safarian, Winona State University
Glenn Shephard, San Jose State University
Jung P. Shim, Mississippi State University
Meenu Singh, Murray State University
Randy Smith, University of Virginia

James T. C. Teng, University of South Carolina
John VanGigch, California State University at Sacramento
David Van Over, University of Idaho
Paul J. A. van Vliet, University of Nebraska at Omaha
B. S. Vijayaraman, University of Akron
Howard Charles Walton, Gettysburg College
Diane B. Walz, University of Texas at San Antonio
Paul R. Watkins, University of Southern California
Randy S. Weinberg, Saint Cloud State University
Jennifer Williams, University of Southern Indiana
Selim Zaim, Sehir University
Steve Zanakis, Florida International University
Fan Zhao, Florida Gulf Coast University
Hamed Majidi Zolbanin, Ball State University

Several individuals contributed material to the text or the supporting material. For this new edition, assistance from the following students and colleagues is gratefully acknowledged: Behrooz Davazdahemami, Bhavana Baheti, Varnika Gottipati, and Chakradhar Pathi (all of Oklahoma State University). Prof. Rick Wilson contributed some examples and new exercise questions for Chapter 8. Prof. Pankush Kalgotra (Auburn University) contributed the new streaming analytics tutorial in Chapter 9. Other contributors of materials for specific application stories are identified as sources in the respective sections. Susan Baskin, Imad Birouty, Sri Raghavan, and Yenny Yang of Teradata provided special help in identifying new TUN content for the book and arranging permissions for the same.

Many other colleagues and students have assisted us in developing previous editions or the recent edition of the companion book from which some of the content has been adapted in this revision. Some of that content is still included this edition. Their assistance and contributions are acknowledged as well in chronological order. Dr. Dave Schrader contributed the sports examples used in Chapter 1. These will provide a great introduction to analytics. We also thank INFORMS for their permission to highlight content from *Interfaces*. We also recognize the following individuals for their assistance in developing Previous edition of the book: Pankush Kalgotra, Prasoon Mathur, Rupesh Agarwal, Shubham Singh, Nan Liang, Jacob Pearson, Kinsey Clemmer, and Evan Murlette (all of Oklahoma State University). Their help for BI 4e is gratefully acknowledged. The Teradata Aster team, especially Mark Ott, provided the material for the opening vignette for Chapter 9. Dr. Brian LeClaire, CIO of Humana Corporation led with contributions of several real-life healthcare case studies developed by his team at Humana. Abhishek Rathi of vCreaTek contributed his vision of analytics in the retail industry. In addition, the following former PhD students and research colleagues of ours have provided content or advice and support for the book in many direct and indirect ways: Asil Oztekin, University of Massachusetts-Lowell; Enes Eryarsoy, Sehir University; Hamed Majidi Zolbanin, Ball State University; Amir Hassan Zadeh, Wright State University; Supavich (Fone) Pengnate, North Dakota State University; Christie Fuller, Boise State University; Daniel Asamoah, Wright State University; Selim Zaim, Istanbul Technical University; and Nihat Kasap, Sabanci University. Peter Horner, editor of *OR/MS Today*, allowed us to summarize new application stories from *OR/MS Today* and *Analytics Magazine*. We also thank INFORMS for their permission to highlight content from *Interfaces*. Assistance from Natraj Ponna, Daniel Asamoah, Amir Hassan-Zadeh, Kartik Dasika, and Angie Jungermann (all of Oklahoma State University) is gratefully acknowledged for DSS 10th edition. We also acknowledge Jongswas Chongwatpol (NIDA, Thailand) for the material on SIMIO software, and Kazim Topuz (University of Tulsa) for his contributions to the Bayesian networks section in

Chapter 5. For other previous editions, we acknowledge the contributions of Dave King (a technology consultant and former executive at JDA Software Group, Inc.) and Jerry Wagner (University of Nebraska–Omaha). Major contributors for earlier editions include Mike Goul (Arizona State University) and Leila A. Halawi (Bethune-Cookman College), who provided material for the chapter on data warehousing; Christy Cheung (Hong Kong Baptist University), who contributed to the chapter on knowledge management; Linda Lai (Macau Polytechnic University of China); Lou Frenzel, an independent consultant whose books *Crash Course in Artificial Intelligence and Expert Systems* and *Understanding of Expert Systems* (both published by Howard W. Sams, New York, 1987) provided material for the early editions; Larry Medsker (American University), who contributed substantial material on neural networks; and Richard V. McCarthy (Quinnipiac University), who performed major revisions in the seventh edition.

Previous editions of the book have also benefited greatly from the efforts of many individuals who contributed advice and interesting material (such as problems), gave feedback on material, or helped with class testing. These include Warren Briggs (Suffolk University), Frank DeBalough (University of Southern California), Mei-Ting Cheung (University of Hong Kong), Alan Dennis (Indiana University), George Easton (San Diego State University), Janet Fisher (California State University, Los Angeles), David Friend (Pilot Software, Inc.), the late Paul Gray (Claremont Graduate School), Mike Henry (OSU), Dustin Huntington (Exsys, Inc.), Subramanian Rama Iyer (Oklahoma State University), Elena Karahanna (The University of Georgia), Mike McAulliffe (The University of Georgia), Chad Peterson (The University of Georgia), Neil Rabjohn (York University), Jim Ragusa (University of Central Florida), Alan Rowe (University of Southern California), Steve Ruth (George Mason University), Linus Schrage (University of Chicago), Antonie Stam (University of Missouri), Late Ron Swift (NCR Corp.), Merril Warkentin (then at Northeastern University), Paul Watkins (The University of Southern California), Ben Mortagy (Claremont Graduate School of Management), Dan Walsh (Bellcore), Richard Watson (The University of Georgia), and the many other instructors and students who have provided feedback.

Several vendors cooperated by providing development and/or demonstration software: Dan Fylstra of Frontline Systems, Gregory Piatetsky-Shapiro of **KDNuggets.com**, Logic Programming Associates (UK), Gary Lynn of NeuroDimension Inc. (Gainesville, Florida), Palisade Software (Newfield, New York), Jerry Wagner of Planners Lab (Omaha, Nebraska), Promised Land Technologies (New Haven, Connecticut), Salford Systems (La Jolla, California), Gary Miner of StatSoft, Inc. (Tulsa, Oklahoma), Ward Systems Group, Inc. (Frederick, Maryland), Idea Fisher Systems, Inc. (Irving, California), and Wordtech Systems (Orinda, California).

Special thanks to the Teradata University Network and especially to Hugh Watson, Michael Goul, and Susan Baskin, Program Director, for their encouragement to tie this book with TUN and for providing useful material for the book.

Many individuals helped us with administrative matters and editing, proofreading, and preparation. The project began with Jack Repcheck (a former Macmillan editor), who initiated this project with the support of Hank Lucas (New York University). Jon Outland assisted with the supplements.

Finally, the Pearson team is to be commended: Executive Editor Samantha Lewis who orchestrated this project; the copyeditors; and the production team, Faraz Sharique Ali at Pearson, and Gowthaman and staff at Integra Software Services, who transformed the manuscript into a book.

We would like to thank all these individuals and corporations. Without their help, the creation of this book would not have been possible. We want to specifically acknowledge the contributions of previous coauthors Janine Aronson, David King, and T. P. Liang, whose original contributions constitute significant components of the book.

R.S.

D.D.

E.T.

Note that Web site URLs are dynamic. As this book went to press, we verified that all the cited Web sites were active and valid. Web sites to which we refer in the text sometimes change or are discontinued because companies change names, are bought or sold, merge, or fail. Sometimes Web sites are down for maintenance, repair, or redesign. Most organizations have dropped the initial "www" designation for their sites, but some still use it. If you have a problem connecting to a Web site that we mention, please be patient and simply run a Web search to try to identify the new site. Most times, the new site can be found quickly. Some sites also require a free registration before allowing you to see the content. We apologize in advance for this inconvenience.

ABOUT THE AUTHORS

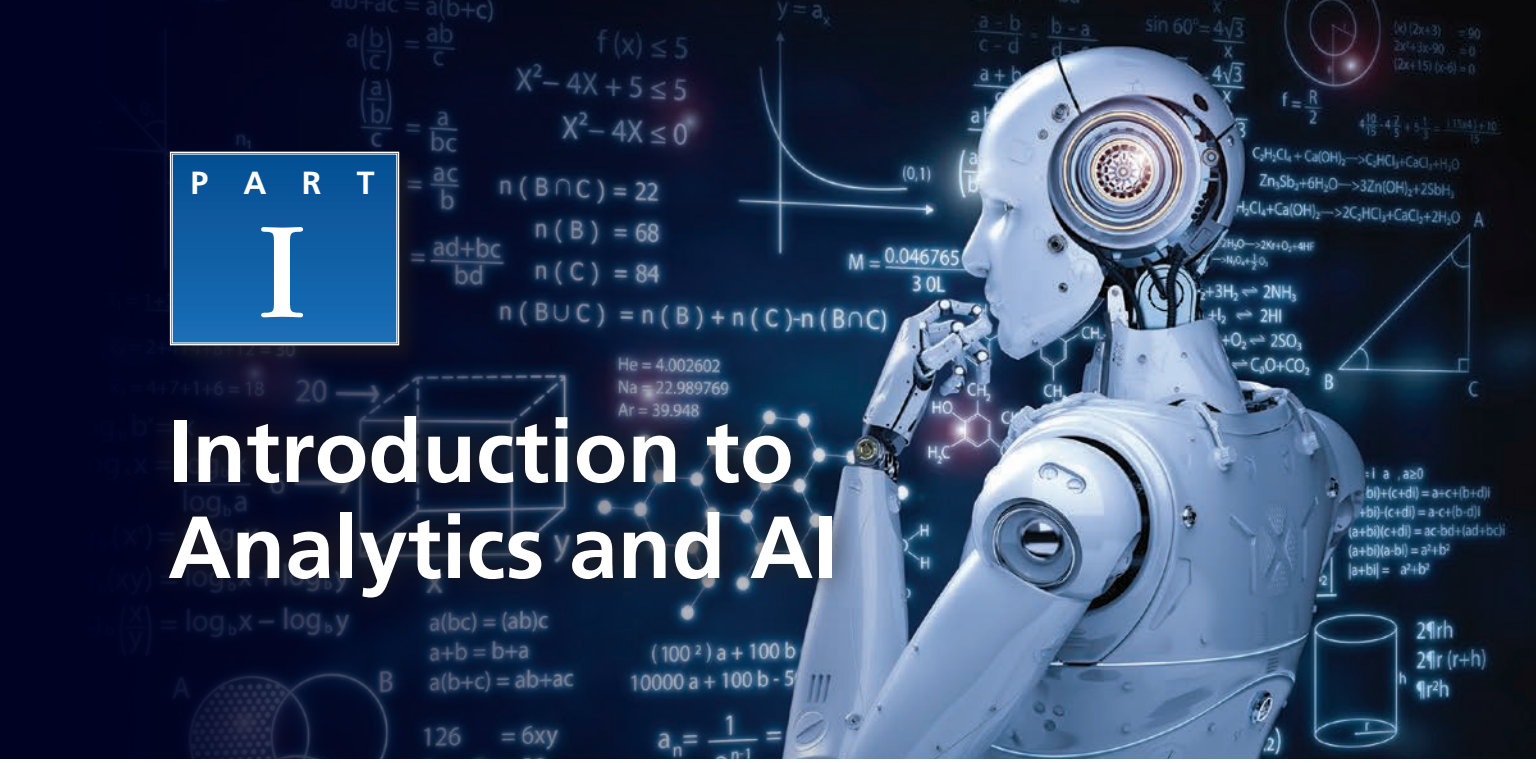
Ramesh Sharda (M.B.A., Ph.D., University of Wisconsin–Madison) is the Vice Dean for Research and Graduate Programs, Watson/ConocoPhillips Chair and a Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University. His research has been published in major journals in management science and information systems including *Management Science, Operations Research, Information Systems Research, Decision Support Systems, Decision Sciences Journal, EJIS, JMIS, Interfaces, INFORMS Journal on Computing, ACM Data Base*, and many others. He is a member of the editorial boards of journals such as the *Decision Support Systems, Decision Sciences*, and *ACM Database*. He has worked on many sponsored research projects with government and industry, and has also served as consultants to many organizations. He also serves as the Faculty Director of Teradata University Network. He received the 2013 INFORMS Computing Society HG Lifetime Service Award, and was inducted into Oklahoma Higher Education Hall of Fame in 2016. He is a Fellow of INFORMS.

Dursun Delen (Ph.D., Oklahoma State University) is the Spears and Patterson Chairs in Business Analytics, Director of Research for the Center for Health Systems Innovation, and Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University (OSU). Prior to his academic career, he worked for a privately owned research and consultancy company, Knowledge Based Systems Inc., in College Station, Texas, as a research scientist for five years, during which he led a number of decision support and other information systems–related research projects funded by federal agencies such as DoD, NASA, NIST, and DOE. Dr. Delen's research has appeared in major journals including *Decision Sciences, Decision Support Systems, Communications of the ACM, Computers and Operations Research, Computers in Industry, Journal of Production Operations Management, Journal of American Medical Informatics Association, Artificial Intelligence in Medicine, Expert Systems with Applications*, among others. He has published eight books/textbooks and more than 100 peer-reviewed journal articles. He is often invited to national and international conferences for keynote addresses on topics related to business analytics, Big Data, data/text mining, business intelligence, decision support systems, and knowledge management. He served as the general co-chair for the 4th International Conference on Network Computing and Advanced Information Management (September 2–4, 2008, in Seoul, South Korea) and regularly serves as chair on tracks and mini-tracks at various business analytics and information systems conferences. He is the co-editor-in-chief for the *Journal of Business Analytics*, the area editor for Big Data and Business Analytics on the Journal of Business Research, and also serves as chief editor, senior editor, associate editor, and editorial board member on more than a dozen other journals. His consultancy, research, and teaching interests are in business analytics, data and text mining, health analytics, decision support systems, knowledge management, systems analysis and design, and enterprise modeling.

Efraim Turban (M.B.A., Ph.D., University of California, Berkeley) is a visiting scholar at the Pacific Institute for Information System Management, University of Hawaii. Prior to this, he was on the staff of several universities, including City University of Hong Kong; Lehigh University; Florida International University; California State University, Long

Beach; Eastern Illinois University; and the University of Southern California. Dr. Turban is the author of more than 110 refereed papers published in leading journals, such as *Management Science, MIS Quarterly*, and *Decision Support Systems*. He is also the author of 22 books, including *Electronic Commerce: A Managerial Perspective* and *Information Technology for Management*. He is also a consultant to major corporations worldwide. Dr. Turban's current areas of interest are Web-based decision support systems, digital commerce, and applied artificial intelligence.

This page is intentionally left blank



Introduction to Analytics and AI

CHAPTER

1

Overview of Business Intelligence, Analytics, Data Science, and Artificial Intelligence: Systems for Decision Support

LEARNING OBJECTIVES

- Understand the need for computerized support of managerial decision making
- Understand the development of systems for providing decision-making support
- Recognize the evolution of such computerized support to the current state of analytics/data science and artificial intelligence
- Describe the business intelligence (BI) methodology and concepts
- Understand the different types of analytics and review selected applications
- Understand the basic concepts of artificial intelligence (AI) and see selected applications
- Understand the analytics ecosystem to identify various key players and career opportunities

The business environment (climate) is constantly changing, and it is becoming more and more complex. Organizations, both private and public, are under pressures that force them to respond quickly to changing conditions and to be innovative in the way they operate. Such activities require organizations to be agile and to make frequent and quick strategic, tactical, and operational decisions, some of which are very complex. Making such decisions may require considerable amounts of relevant data, information, and knowledge. Processing these in the framework of the needed decisions must be done quickly, frequently in real time, and usually requires some computerized support. As technologies are evolving, many decisions are being automated, leading to a major impact on knowledge work and workers in many ways.

This book is about using business analytics and artificial intelligence (AI) as a computerized support portfolio for managerial decision making. It concentrates on the

theoretical and conceptual foundations of decision support as well as on the commercial tools and techniques that are available. The book presents the fundamentals of the techniques and the manner in which these systems are constructed and used. We follow an EEE (*exposure, experience*, and *exploration*) approach to introducing these topics. The book primarily provides exposure to various analytics/AI techniques and their applications. The idea is that students will be inspired to learn from how various organizations have employed these technologies to make decisions or to gain a competitive edge. We believe that such exposure to what is being accomplished with analytics and that how it can be achieved is the key component of learning about analytics. In describing the techniques, we also give examples of specific software tools that can be used for developing such applications. However, the book is not limited to any one software tool, so students can experience these techniques using any number of available software tools. We hope that this *exposure* and *experience* enable and motivate readers to *explore* the potential of these techniques in their own domain. To facilitate such exploration, we include exercises that direct the reader to Teradata University Network (TUN) and other sites that include team-oriented exercises where appropriate. In our own teaching experience, projects undertaken in the class facilitate such exploration after students have been exposed to the myriad of applications and concepts in the book and they have experienced specific software introduced by the professor.

This introductory chapter provides an introduction to analytics and artificial intelligence as well as an overview of the book. The chapter has the following sections:

1.1 OPENING VIGNETTE: How Intelligent Systems Work for KONE Elevators and Escalators Company

KONE is a global industrial company (based in Finland) that manufactures mostly elevators and escalators and also services over 1.1 million elevators, escalators, and related equipment in several countries. The company employs over 50,000 people.

THE PROBLEM

Over 1 billion people use the elevators and escalators manufactured and serviced by KONE every day. If equipment does not work properly, people may be late to work, cannot get home in time, and may miss important meetings and events. So, KONE's objective is to minimize the downtime and users' suffering.

The company has over 20,000 technicians who are dispatched to deal with the elevators anytime a problem occurs. As buildings are getting higher (the trend in many places), more people are using elevators, and there is more pressure on elevators to handle the growing amount of traffic. KONE faced the responsibility to serve users smoothly and safely.

THE SOLUTION

KONE decided to use IBM Watson IoT Cloud platform. As we will see in Chapter 6, IBM installed cognitive abilities in buildings that make it possible to recognize situations and behavior of both people and equipment. The Internet of Things (IoT), as we will see in Chapter 13, is a platform that can connect millions of "things" together and to a central command that can manipulate the connected things. Also, the IoT connects sensors that are attached to KONE's elevators and escalators. The sensors collect information and data about the elevators (such as noise level) and other equipment in real time. Then, the IoT transfers to information centers via the collected data "cloud." There, analytic systems (IBM Advanced Analytic Engine) and AI process the collected data and predict things such as potential failures. The systems also identify the likely causes of problems and suggest potential remedies. Note the predictive power of IBM Watson Analytics (using machine learning, an AI technology described in Chapters 4–6) for finding problems before they occur.

The KONE system collects a significant amount of data that are analyzed for other purposes so that future design of equipment can be improved. This is because Watson Analytics offers a convenient environment for communication of and collaboration around the data. In addition, the analysis suggests how to optimize buildings and equipment operations. Finally, KONE and its customers can get insights regarding the financial aspects of managing the elevators.

KONE also integrates the Watson capabilities with Salesforce's service tools (Service Cloud Lightning and Field Service Lightning). This combination helps KONE to immediately respond to emergencies or soon-to-occur failures as quickly as possible, dispatching some of its 20,000 technicians to the problems' sites. Salesforce also provides superb customer relationship management (CRM). The people–machine communication, query, and collaboration in the system are in a natural language (an AI capability of Watson Analytics; see Chapter 6). Note that IBM Watson analytics includes two types of analytics: *predictive*, which predicts when failures may occur, and *prescriptive*, which recommends actions (e.g., preventive maintenance).

THE RESULTS

KONE has minimized downtime and shortened the repair time. Obviously, elevators/escalators users are much happier if they do not have problems because of equipment downtime, so they enjoy trouble-free rides. The prediction of "soon-to-happen" can save many problems for the equipment owners. The owners can also optimize the schedule of their own employees (e.g., cleaners and maintenance workers). All in all, the decision makers at both KONE and the buildings can make informed and better decisions. Some day in the future, robots may perform maintenance and repairs of elevators and escalators.

Note: This case is a sample of IBM Watson's success using its cognitive buildings capability. To learn more, we suggest you view the following YouTube videos: (1) **youtube.com/watch?v=6UPJHyiJft0** (1:31 min.) (2017); (2) **youtube.com/watch?v=EVbd3ejEXus** (2:49 min.) (2017).

Sources: Compiled from J. Fernandez. (2017, April). "A Billion People a Day. Millions of Elevators. No Room for Downtime." IBM developer Works Blog. **developer.ibm.com/dwblog/2017/kone-watson-video/** (accessed September 2018); H. Srikanthan. "KONE Improves 'People Flow' in 1.1 Million Elevators with IBM Watson IoT." Generis. **https://generisgp.com/2018/01/08/ibm-case-study-kone-corp/** (accessed September 2018); L. Slowey. (2017, February 16). "Look Who's Talking: KONE Makes Elevator Services Truly Intelligent with Watson IoT." IBM Internet of Things Blog. **ibm.com/blogs/internet-of-things/kone/** (accessed September 2018).

QUESTIONS FOR THE OPENING VIGNETTE

1. It is said that KONE is embedding intelligence across its supply chain and enables smarter buildings. Explain.
2. Describe the role of IoT in this case.
3. What makes IBM Watson a necessity in this case?
4. Check IBM Advanced Analytics. What tools were included that relate to this case?
5. Check IBM cognitive buildings. How do they relate to this case?

WHAT CAN WE LEARN FROM THIS VIGNETTE?

Today, intelligent technologies can embark on large-scale complex projects when they include AI combined with IoT. The capabilities of integrated intelligent platforms, such as IBM Watson, make it possible to solve problems that were economically and technologically unsolvable just a few years ago. The case introduces the reader to several of the technologies, including advanced analytics, sensors, IoT, and AI that are covered in this book. The case also points to the use of "cloud." The cloud is used to centrally process large amounts of information using analytics and AI algorithms, involving "things" in different locations. This vignette also introduces us to two major types of analytics: predictive analytics (Chapters 4–6) and prescriptive analytics (Chapter 8).

Several AI technologies are discussed: machine learning, natural language processing, computer vision, and prescriptive analysis.

The case is an example of *augmented intelligence* in which people and machines work together. The case illustrates the benefits to the vendor, the implementing companies, and their employees and to the users of the elevators and escalators.

1.2 CHANGING BUSINESS ENVIRONMENTS AND EVOLVING NEEDS FOR DECISION SUPPORT AND ANALYTICS

Decision making is one of the most important activities in organizations of all kind—probably the most important one. Decision making leads to the success or failure of organizations and how well they perform. Making decisions is getting difficult due to internal and external factors. The rewards of making appropriate decisions can be very high and so can the loss of inappropriate ones.

Unfortunately, it is not simple to make decisions. To begin with, there are several types of decisions, each of which requires a different decision-making approach. For example, De Smet et al. (2017) of McKinsey & Company management consultants classify organizational decision into the following four groups:

- Big-bet, high-risk decisions.
- Cross-cutting decisions, which are repetitive but high risk that require group work (Chapter 11).
- Ad hoc decisions that arise episodically.
- Delegated decisions to individuals or small groups.

Therefore, it is necessary first to understand the nature of decision making. For a comprehensive discussion, see (De Smet et al. 2017).

Modern business is full of uncertainties and rapid changes. To deal with these, organizational decision makers need to deal with ever-increasing and changing data. This book is about the technologies that can assist decision makers in their jobs.

Decision-Making Process

For years, managers considered decision making purely an art—a talent acquired over a long period through experience (i.e., learning by trial and error) and by using intuition. Management was considered an art because a variety of individual styles could be used in approaching and successfully solving the same types of managerial problems. These styles were often based on creativity, judgment, intuition, and experience rather than on systematic quantitative methods grounded in a scientific approach. However, recent research suggests that companies with top managers who are more focused on persistent work tend to outperform those with leaders whose main strengths are interpersonal communication skills. It is more important to emphasize methodical, thoughtful, analytical decision making rather than flashiness and interpersonal communication skills.

Managers usually make decisions by following a four-step process (we learn more about these in the next section):

1. Define the problem (i.e., a decision situation that may deal with some difficulty or with an opportunity).
2. Construct a model that describes the real-world problem.
3. Identify possible solutions to the modeled problem and evaluate the solutions.
4. Compare, choose, and recommend a potential solution to the problem.

A more detailed process is offered by Quain (2018), who suggests the following steps:

1. Understand the decision you have to make.
2. Collect all the information.
3. Identify the alternatives.
4. Evaluate the pros and cons.
5. Select the best alternative.
6. Make the decision.
7. Evaluate the impact of your decision.

We will return to this process in Section 1.3.

The Influence of the External and Internal Environments on the Process

To follow these decision-making processes, one must make sure that sufficient alternative solutions, including good ones, are being considered, that the consequences of using these alternatives can be reasonably predicted, and that comparisons are done properly. However, rapid changes in internal and external environments make such an evaluation process difficult for the following reasons:

- Technology, information systems, advanced search engines, and globalization result in more and more alternatives from which to choose.
- Government regulations and the need for compliance, political instability and terrorism, competition, and changing consumer demands produce more uncertainty, making it more difficult to predict consequences and the future.
  - **Political factors.** Major decisions may be influenced by both external and internal politics. An example is the 2018 trade war on tariffs.
  - **Economic factors.** These range from competition to the genera and state of the economy. These factors, both in the short and long run, need to be considered.

- **Sociological and psychological factors regarding employees and customers.** These need to be considered when changes are being made.
- **Environment factors.** The impact on the physical environment must be assessed in many decision-making situations.

Other factors include the need to make rapid decisions, the frequent and unpredictable changes that make trial-and-error learning difficult, and the potential costs of making mistakes that may be large.

These environments are growing more complex every day. Therefore, making decisions today is indeed a complex task. For further discussion, see Charles (2018). For how to make effective decisions under uncertainty and pressure, see Zane (2016).

Because of these trends and changes, it is nearly impossible to rely on a trial-and-error approach to management. Managers must be more sophisticated; they must use the new tools and techniques of their fields. Most of those tools and techniques are discussed in this book. Using them to support decision making can be extremely rewarding in making effective decisions. Further, many tools that are evolving impact even the very existence of several decision-making tasks that are being automated. This impacts future demand for knowledge workers and begs many legal and societal impact questions.

Data and Its Analysis in Decision Making

We will see several times in this book how an entire industry can employ analytics to develop reports on what is happening, predict what is likely to happen, and then make decisions to make the best use of the situation at hand. These steps require an organization to collect and analyze vast stores of data. In general, the amount of data doubles every two years. From traditional uses in payroll and bookkeeping functions, computerized systems are now used for complex managerial areas ranging from the design and management of automated factories to the application of analytical methods for the evaluation of proposed mergers and acquisitions. Nearly all executives know that information technology is vital to their business and extensively use these technologies.

Computer applications have moved from transaction-processing and monitoring activities to problem analysis and solution applications, and much of the activity is done with cloud-based technologies, in many cases accessed through mobile devices. Analytics and BI tools such as data warehousing, data mining, online analytical processing (OLAP), dashboards, and the use of cloud-based systems for decision support are the cornerstones of today's modern management. Managers must have high-speed, networked information systems (wired or wireless) to assist them with their most important task: making decisions. In many cases, such decisions are routinely being fully automated (see Chapter 2), eliminating the need for any managerial intervention.

Technologies for Data Analysis and Decision Support

Besides the obvious growth in hardware, software, and network capacities, some developments have clearly contributed to facilitating the growth of decision support and analytics technologies in a number of ways:

- **Group communication and collaboration.** Many decisions are made today by groups whose members may be in different locations. Groups can collaborate and communicate readily by using collaboration tools as well as the ubiquitous smartphones. Collaboration is especially important along the supply chain, where partners—all the way from vendors to customers—must share information. Assembling a group of decision makers, especially experts, in one place can be

costly. Information systems can improve the collaboration process of a group and enable its members to be at different locations (saving travel costs). More critically, such supply chain collaboration permits manufacturers to know about the changing patterns of demand in near real time and thus react to marketplace changes faster. For a comprehensive coverage and the impact of AI, see Chapters 2, 10, and 14.

- **Improved data management.** Many decisions involve complex computations. Data for these can be stored in different databases anywhere in the organization and even possibly outside the organization. The data may include text, sound, graphics, and video, and these can be in different languages. Many times it is necessary to transmit data quickly from distant locations. Systems today can search, store, and transmit needed data quickly, economically, securely, and transparently. See Chapters 3 and 9 and the online chapter for details.
- **Managing giant data warehouses and Big Data.** Large data warehouses (DWs), like the ones operated by Walmart, contain huge amounts of data. Special methods, including parallel computing and Hadoop/Spark, are available to organize, search, and mine the data. The costs related to data storage and mining are declining rapidly. Technologies that fall under the broad category of Big Data have enabled massive data coming from a variety of sources and in many different forms, which allows a very different view of organizational performance that was not possible in the past. See Chapter 9 for details.
- **Analytical support.** With more data and analysis technologies, more alternatives can be evaluated, forecasts can be improved, risk analysis can be performed quickly, and the views of experts (some of whom may be in remote locations) can be collected quickly and at a reduced cost. Expertise can even be derived directly from analytical systems. With such tools, decision makers can perform complex simulations, check many possible scenarios, and assess diverse impacts quickly and economically.This, of course, is the focus of several chapters in the book. See Chapters 4–7.
- **Overcoming cognitive limits in processing and storing information.** The human mind has only a limited ability to process and store information. People sometimes find it difficult to recall and use information in an error-free fashion due to their cognitive limits. The term *cognitive limits* indicates that an individual's problem-solving capability is limited when a wide range of diverse information and knowledge is required. Computerized systems enable people to overcome their cognitive limits by quickly accessing and processing vast amounts of stored information. One way to overcome humans' cognitive limitations is to use AI support. For coverage of cognitive aspects, see Chapter 6.
- **Knowledge management.** Organizations have gathered vast stores of information about their own operations, customers, internal procedures, employee interactions, and so forth through the unstructured and structured communications taking place among various stakeholders. Knowledge management systems (KMS) have become sources of formal and informal support for decision making to managers, although sometimes they may not even be called *KMS*. Technologies such as text analytics and IBM Watson are making it possible to generate value from such knowledge stores. (See Chapters 6 and 12 for details.
- **Anywhere, anytime support.** Using wireless technology, managers can access information anytime and from any place, analyze and interpret it, and communicate with those using it. This perhaps is the biggest change that has occurred in the last few years. The speed at which information needs to be processed and converted into decisions has truly changed expectations for both consumers and businesses. These and other capabilities have been driving the use of computerized decision support since the late 1960s, especially since the mid-1990s. The growth of mobile technologies, social media platforms, and analytical tools has enabled a different level of information systems (IS) to support managers. This growth in providing

data-driven support for any decision extends not just to managers but also to consumers. We will first study an overview of technologies that have been broadly referred to as BI. From there we will broaden our horizons to introduce various types of analytics.

- **Innovation and artificial intelligence.** Because of the complexities in the decision-making process discussed earlier and the environment surrounding the process, a more innovative approach is frequently need. A major facilitation of innovation is provided by AI. Almost every step in the decision-making process can be influenced by AI. AI is also integrated with analytics, creating synergy in making decisions (Section 1.8).

SECTION 1.2 REVIEW QUESTIONS

1. Why is it difficult to make organizational decisions?
2. Describe the major steps in the decision-making process.
3. Describe the major external environments that can impact decision making.
4. What are some of the key system-oriented trends that have fostered IS-supported decision making to a new level?
5. List some capabilities of information technologies that can facilitate managerial decision making.

1.3 DECISION-MAKING PROCESSES AND COMPUTERIZED DECISION SUPPORT FRAMEWORK

In this section, we focus on some classical decision-making fundamentals and in more detail on the decision-making process. These two concepts will help us ground much of what we will learn in terms of analytics, data science, and artificial intelligence.

Decision making is a process of choosing among two or more alternative courses of action for the purpose of attaining one or more goals. According to Simon (1977), managerial decision making is synonymous with the entire management process. Consider the important managerial function of planning. Planning involves a series of decisions: What should be done? When? Where? Why? How? By whom? Managers set goals, or plan; hence, planning implies decision making. Other managerial functions, such as organizing and controlling, also involve decision making.

Simon's Process: Intelligence, Design, and Choice

It is advisable to follow a systematic decision-making process. Simon (1977) said that this involves three major phases: intelligence, design, and choice. He later added a fourth phase: implementation. Monitoring can be considered a fifth phase—a form of feedback. However, we view monitoring as the *intelligence phase* applied to the *implementation phase*. Simon's model is the most concise and yet complete characterization of rational decision making. A conceptual picture of the decision-making process is shown in Figure 1.1. It is also illustrated as a decision support approach using modeling.

There is a continuous flow of activity from intelligence to design to choice (see the solid lines in Figure 1.1), but at any phase, there may be a return to a previous phase (feedback). Modeling is an essential part of this process. The seemingly chaotic nature of following a haphazard path from problem discovery to solution via decision making can be explained by these feedback loops.

The decision-making process starts with the **intelligence phase**; in this phase, the decision maker examines reality and identifies and defines the problem. *Problem ownership* is established as well. In the **design phase**, a model that represents the system is constructed. This is done by making assumptions that simplify reality and by writing down

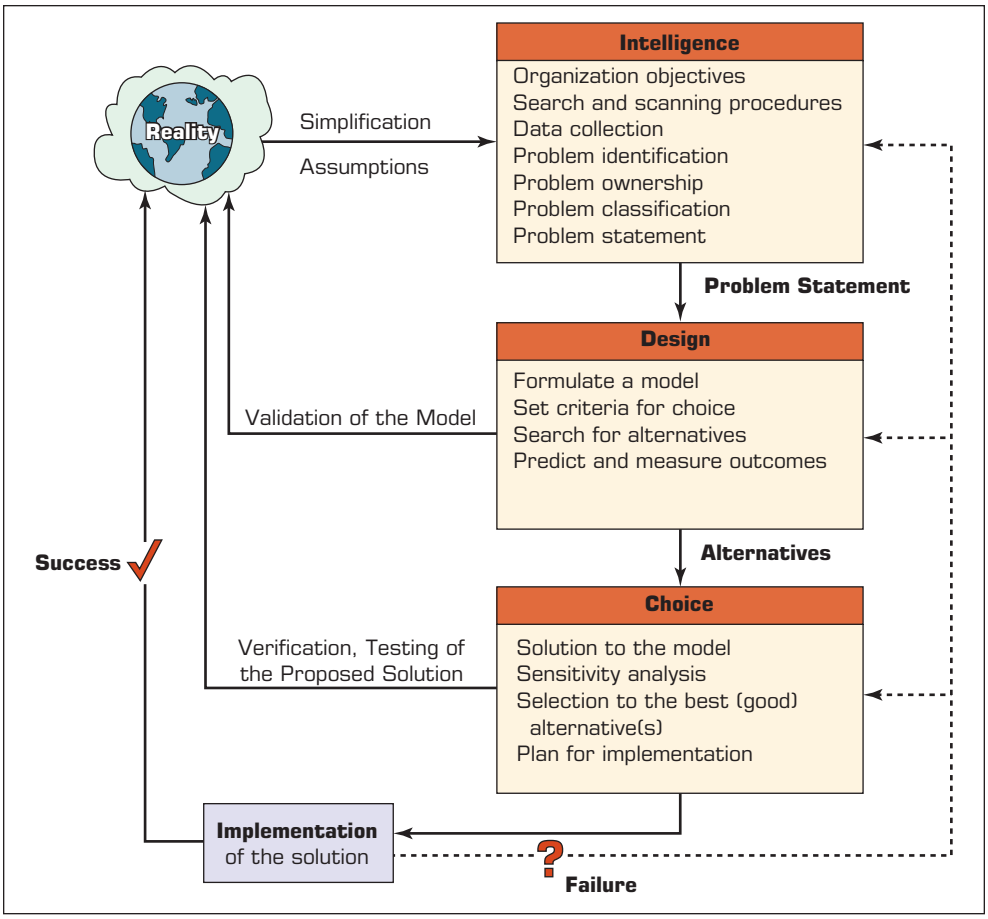


FIGURE 1.1 **The Decision-Making/Modeling Process.**

the relationships among all the variables. The model is then validated, and criteria are determined in a principle of choice for evaluation of the alternative courses of action that are identified. Often, the process of model development identifies alternative solutions and vice versa.

The **choice phase** includes the selection of a proposed solution to the model (not necessarily to the problem it represents). This solution is tested to determine its viability. When the proposed solution seems reasonable, we are ready for the last phase: implementation of the decision (not necessarily of a system). Successful implementation results in solving the real problem. Failure leads to a return to an earlier phase of the process. In fact, we can return to an earlier phase during any of the latter three phases. The decision-making situations described in the opening vignette follow Simon's four-phase model, as do almost all other decision-making situations.

The Intelligence Phase: Problem (or Opportunity) Identification

The intelligence phase begins with the identification of organizational goals and objectives related to an issue of concern (e.g., inventory management, job selection, lack of or incorrect Web presence) and determination of whether they are being met. Problems occur because of dissatisfaction with the status quo. Dissatisfaction is the result of a difference between what people desire (or expect) and what is occurring. In this first phase, a decision maker attempts to determine whether a problem exists, identify its symptoms, determine its magnitude, and

explicitly define it. Often, what is described as a problem (e.g., excessive costs) may be only a symptom (i.e., measure) of a problem (e.g., improper inventory levels). Because real-world problems are usually complicated by many interrelated factors, it is sometimes difficult to distinguish between the symptoms and the real problem. New opportunities and problems certainly may be uncovered while investigating the causes of symptoms.

The existence of a problem can be determined by monitoring and analyzing the organization's productivity level. The measurement of productivity and the construction of a model are based on real data. The collection of data and the estimation of future data are among the most difficult steps in the analysis.

ISSUES IN DATA COLLECTION The following are some issues that may arise during data collection and estimation and thus plague decision makers:

- Data are not available. As a result, the model is made with and relies on potentially inaccurate estimates.
- Obtaining data may be expensive.
- Data may not be accurate or precise enough.
- Data estimation is often subjective.
- Data may be insecure.
- Important data that influence the results may be qualitative (soft).
- There may be too many data (i.e., information overload).
- Outcomes (or results) may occur over an extended period. As a result, revenues, expenses, and profits will be recorded at different points in time. To overcome this difficulty, a present-value approach can be used if the results are quantifiable.
- It is assumed that future data will be similar to historical data. If this is not the case, the nature of the change has to be predicted and included in the analysis.

When the preliminary investigation is completed, it is possible to determine whether a problem really exists, where it is located, and how significant it is. A key issue is whether an information system is reporting a problem or only the symptoms of a problem. For example, if reports indicate that sales are down, there is a problem, but the situation, no doubt, is symptomatic of the problem. It is critical to know the real problem. Sometimes it may be a problem of perception, incentive mismatch, or organizational processes rather than a poor decision model.

To illustrate why it is important to identify the problem correctly, we provide a classical example in Application Case 1.1.

Application Case 1.1 Making Elevators Go Faster!

This story has been reported in numerous places and has almost become a classic example to explain the need for problem identification. Ackoff (as cited in Larson, 1987) described the problem of managing complaints about slow elevators in a tall hotel tower. After trying many solutions for reducing the complaint—staggering elevators to go to different floors, adding operators, and so on—the management determined that the real problem was not about the *actual* waiting time but rather the *perceived* waiting time. So the solution was to install full-length mirrors on elevator doors on each floor. As Hesse and Woolsey (1975) put it, "The women would look at themselves in the mirrors and make adjustments, while the men would look at the women, and before they knew it, the elevator was there." By reducing the perceived waiting time, the problem went away. Baker and Cameron (1996)

(*Continued*)

Application Case 1.1 (Continued)

give several other examples of distractions, including lighting and displays, that organizations use to reduce perceived waiting time. If the real problem is identified as *perceived* waiting time, it can make a big difference in the proposed solutions and their costs. For example, full-length mirrors probably cost a whole lot less than adding an elevator!

Sources: Based on J. Baker and M. Cameron. (1996, September). "The Effects of the Service Environment on Affect and Consumer Perception of Waiting Time: An Integrative Review and Research Propositions," *Journal of the Academy of Marketing Science, 24,* pp. 338–349; R. Hesse and G. Woolsey (1975). *Applied Management Science: A Quick and Dirty Approach.* Chicago, IL: SRA Inc; R. C. Larson. (1987, November/December). "Perspectives on Queues: Social Justice and the Psychology of Queuing." *Operations Research, 35*(6), pp. 895–905.

Questions for Case 1.1

1. Why this is an example relevant to decision making?
2. Relate this situation to the intelligence phase of decision making.

PROBLEM CLASSIFICATION Problem classification is the conceptualization of a problem in an attempt to place it in a definable category, possibly leading to a standard solution approach. An important approach classifies problems according to the degree of structuredness evident in them. This ranges from totally structured (i.e., programmed) to totally unstructured (i.e., unprogrammed).

PROBLEM DECOMPOSITION Many complex problems can be divided into subproblems. Solving the simpler subproblems may help in solving a complex problem. Also, seemingly poorly structured problems sometimes have highly structured subproblems. Just as a semistructured problem results when some phases of decision making are structured whereas other phases are unstructured, and when some subproblems of a decision-making problem are structured with others unstructured, the problem itself is semistructured. As a decision support system is developed and the decision maker and development staff learn more about the problem, it gains structure.

PROBLEM OWNERSHIP In the intelligence phase, it is important to establish problem ownership. A problem exists in an organization only if someone or some group takes the responsibility for attacking it and if the organization has the ability to solve it. The assignment of authority to solve the problem is called *problem ownership*. For example, a manager may feel that he or she has a problem because interest rates are too high. Because interest rate levels are determined at the national and international levels and most managers can do nothing about them, high interest rates are the problem of the government, not a problem for a specific company to solve. The problem that companies actually face is how to operate in a high interest-rate environment. For an individual company, the interest rate level should be handled as an uncontrollable (environmental) factor to be predicted.

When problem ownership is not established, either someone is not doing his or her job or the problem at hand has yet to be identified as belonging to anyone. It is then important for someone to either volunteer to own it or assign it to someone.

The intelligence phase ends with a formal problem statement.

The Design Phase

The design phase involves finding or developing and analyzing possible courses of action. These include understanding the problem and testing solutions for feasibility. A model of the decision-making problem is constructed, tested, and validated. Let us first define a model.

MODELS A major characteristic of computerized decision support and many BI tools (notably those of business analytics) is the inclusion of at least one model. The basic idea is to perform the analysis on a model of reality rather than on the real system. A *model* is a simplified representation or abstraction of reality. It is usually simplified because reality is too complex to describe exactly and because much of the complexity is actually irrelevant in solving a specific problem.

Modeling involves conceptualizing a problem and abstracting it to quantitative and/or qualitative form. For a mathematical model, the variables are identified and their mutual relationships are established. Simplifications are made, whenever necessary, through assumptions. For example, a relationship between two variables may be assumed to be linear even though in reality there may be some nonlinear effects. A proper balance between the level of model simplification and the representation of reality must be obtained because of the cost–benefit trade-off. A simpler model leads to lower development costs, easier manipulation, and a faster solution but is less representative of the real problem and can produce inaccurate results. However, a simpler model generally requires fewer data, or the data are aggregated and easier to obtain.

The Choice Phase

Choice is the critical act of decision making. The choice phase is the one in which the actual decision and the commitment to follow a certain course of action are made. The boundary between the design and choice phases is often unclear because certain activities can be performed during both of them and because the decision maker can return frequently from choice activities to design activities (e.g., generate new alternatives while performing an evaluation of existing ones). The choice phase includes the search for, evaluation of, and recommendation of an appropriate solution to a model. A solution to a model is a specific set of values for the decision variables in a selected alternative. Choices can be evaluated as to their viability and profitability.

Each alternative must be evaluated. If an alternative has multiple goals, they must all be examined and balanced against each other. Sensitivity analysis is used to determine the robustness of any given alternative; slight changes in the parameters should ideally lead to slight or no changes in the alternative chosen. What-if analysis is used to explore major changes in the parameters. Goal seeking helps a manager determine values of the decision variables to meet a specific objective. These topics are addressed in Chapter 8.

The Implementation Phase

In *The Prince*, Machiavelli astutely noted some 500 years ago that there was "nothing more difficult to carry out, nor more doubtful of success, nor more dangerous to handle, than to initiate a new order of things." The implementation of a proposed solution to a problem is, in effect, the initiation of a new order of things or the introduction of change. And change must be managed. User expectations must be managed as part of change management.

The definition of *implementation* is somewhat complicated because implementation is a long, involved process with vague boundaries. Simplistically, the **implementation phase** involves putting a recommended solution to work, not necessarily implementing a computer system. Many generic implementation issues, such as resistance to change, degree of support of top management, and user training, are important in dealing with information system–supported decision making. Indeed, many previous technology-related waves (e.g., business process reengineering [BPR] and knowledge management) have faced mixed results mainly because of change management challenges and issues. Management of change is almost an entire discipline in itself, so we recognize its importance and encourage readers to focus on it independently. Implementation also includes

a thorough understanding of project management. The importance of project management goes far beyond analytics, so the last few years have witnessed a major growth in certification programs for project managers. A very popular certification now is the Project Management Professional (PMP). See **pmi.org** for more details.

Implementation must also involve collecting and analyzing data to learn from the previous decisions and improve the next decision. Although analysis of data is usually conducted to identify the problem and/or the solution, analytics should also be employed in the feedback process. This is especially true for any public policy decisions. We need to be sure that the data being used for problem identification is valid. Sometimes people find this out only after the implementation phase.

The decision-making process, though conducted by people, can be improved with computer support, which is introduced next.

The Classical Decision Support System Framework

The early definitions of decision support system (DSS) identified it as a system intended to support managerial decision makers in semistructured and unstructured decision situations. DSS was meant to be an adjunct to decision makers, extending their capabilities but not replacing their judgment. DSS was aimed at decisions that required judgment or at decisions that could not be completely supported by algorithms. Not specifically stated but implied in the early definitions was the notion that the system would be computer based, would operate interactively online, and preferably would have graphical output capabilities, now simplified via browsers and mobile devices.

An early framework for computerized decision support includes several major concepts that are used in forthcoming sections and chapters of this book. Gorry and Scott-Morton created and used this framework in the early 1970s, and the framework then evolved into a new technology called *DSS*.

Gorry and Scott-Morton (1971) proposed a framework that is a 3-by-3 matrix, as shown in Figure 1.2. The two dimensions are the degree of structuredness and the types of control.

DEGREE OF STRUCTUREDNESS The left side of Figure 1.2 is based on Simon's (1977) idea that decision-making processes fall along a continuum that ranges from highly structured (sometimes called *programmed*) to highly unstructured (i.e., *non-programmed*) decisions. Structured processes are routine and typically repetitive problems for which standard solution methods exist. *Unstructured processes* are fuzzy, complex problems for which there are no cut-and-dried solution methods.

An *unstructured problem* is one where the articulation of the problem or the solution approach may be unstructured in itself. In a *structured problem*, the procedures for obtaining the best (or at least a good enough) solution are known. Whether the problem involves finding an appropriate inventory level or choosing an optimal investment strategy, the objectives are clearly defined. Common objectives are cost minimization and profit maximization.

Semistructured problems fall between structured and unstructured problems, having some structured elements and some unstructured elements. Keen and Scott-Morton (1978) mentioned trading bonds, setting marketing budgets for consumer products, and performing capital acquisition analysis as semistructured problems.

TYPES OF CONTROL The second half of the Gorry and Scott-Morton (1971) framework (refer to Figure 1.2) is based on Anthony's (1965) taxonomy, which defines three broad categories that encompass all managerial activities: *strategic planning*, which involves defining long-range goals and policies for resource allocation; *management control*, the

Type of Decision	Type of Control: Operational Control	Type of Control: Managerial Control	Type of Control: Strategic Planning
Structured	1 Monitoring accounts receivable Monitoring accounts payable Placing order entries	2 Analyzing budget Forecasting short-term Reporting on personnel Making or buying	3 Managing finances Monitoring investment portfolio Locating warehouse Monitoring distribution systems
Semistructured	4 Scheduling production Controlling inventory	5 Evaluating credit Preparing budget Laying out plant Scheduling project Designing reward system Categorizing inventory	6 Building a new plant Planning mergers and acquisitions Planning new products Planning compensation Providing quality assurance Establishing human resources policies Planning inventory
Unstructured	7 Buying software Approving loans Operating a help desk Selecting a cover for a magazine	8 Negotiating Recruiting an executive Buying hardware Lobbying	9 Planning research and development Developing new technologies Planning social responsibility

FIGURE 1.2 **Decision Support Frameworks.**

acquisition and efficient use of resources in the accomplishment of organizational goals; and *operational control*, the efficient and effective execution of specific tasks.

THE DECISION SUPPORT MATRIX Anthony's (1965) and Simon's (1977) taxonomies are combined in the nine-cell decision support matrix shown in Figure 1.2. The initial purpose of this matrix was to suggest different types of computerized support to different cells in the matrix. Gorry and Scott-Morton (1971) suggested, for example, that for making *semistructured decisions* and *unstructured decisions*, conventional management information systems (MIS) and management science (MS) tools are insufficient. Human intellect and a different approach to computer technologies are necessary. They proposed the use of a supportive information system, which they called a *DSS*.

Note that the more structured and operational control-oriented tasks (such as those in cells 1, 2, and 4 of Figure 1.2) are usually performed by lower-level managers, whereas the tasks in cells 6, 8, and 9 are the responsibility of top executives or highly trained specialists.

COMPUTER SUPPORT FOR STRUCTURED DECISIONS Since the 1960s, computers have historically supported structured and some semistructured decisions, especially those that involve operational and managerial control. Operational and managerial control decisions are made in all functional areas, especially in finance and production (i.e., operations) management.

Structured problems, which are encountered repeatedly, have a high level of structure, as their name suggests. It is therefore possible to abstract, analyze, and classify them into specific categories. For example, a make-or-buy decision is one category. Other examples of categories are capital budgeting, allocation of resources, distribution, procurement, planning, and inventory control decisions. For each category of decision, an easy-to-apply prescribed model and solution approach have been developed, generally as quantitative formulas. Therefore, it is possible to use a *scientific approach* for automating portions of managerial decision making. Solutions to many structured problems can be fully automated (see Chapters 2 and 12).

COMPUTER SUPPORT FOR UNSTRUCTURED DECISIONS Unstructured problems can be only partially supported by standard computerized quantitative methods. It is usually necessary to develop customized solutions. However, such solutions may benefit from data and information generated from corporate or external data sources. Intuition and judgment may play a large role in these types of decisions, as may computerized communication and collaboration technologies, as well as cognitive computing (Chapter 6) and deep learning (Chapter 5).

COMPUTER SUPPORT FOR SEMISTRUCTURED PROBLEMS Solving semistructured problems may involve a combination of standard solution procedures and human judgment. Management science can provide models for the portion of a decision-making problem that is structured. For the unstructured portion, a DSS can improve the quality of the information on which the decision is based by providing, for example, not only a single solution, but also a range of alternative solutions along with their potential impacts. These capabilities help managers to better understand the nature of problems and, thus, to make better decisions.

DECISION SUPPORT SYSTEM: CAPABILITIES The early definitions of DSS identified it as a system intended to support managerial decision makers in semistructured and unstructured decision situations. DSS was meant to be an adjunct to decision makers, extending their capabilities but not replacing their judgment. It was aimed at decisions that required judgment or at decisions that could not be completely supported by algorithms. Not specifically stated but implied in the early definitions was the notion that the system would be computer based, would operate interactively online, and preferably would have graphical output capabilities, now simplified via browsers and mobile devices.

A DSS Application

A DSS is typically built to support the solution of a certain problem or to evaluate an opportunity. This is a key difference between DSS and BI applications. In a very strict sense, **business intelligence (BI)** systems monitor situations and identify problems and/or opportunities using analytic methods. Reporting plays a major role in BI; the user generally must identify whether a particular situation warrants attention and then can apply analytical methods. Again, although models and data access (generally through a data warehouse) are included in BI, a DSS may have its own databases and is developed to solve a specific problem or set of problems and are therefore called DSS applications.

Formally, a DSS is an approach (or methodology) for supporting decision making. It uses an interactive, flexible, adaptable computer-based information system (CBIS) especially developed for supporting the solution to a specific unstructured management problem. It uses data, provides an easy user interface, and can incorporate the decision maker's own insights. In addition, a DSS includes models and is developed (possibly by

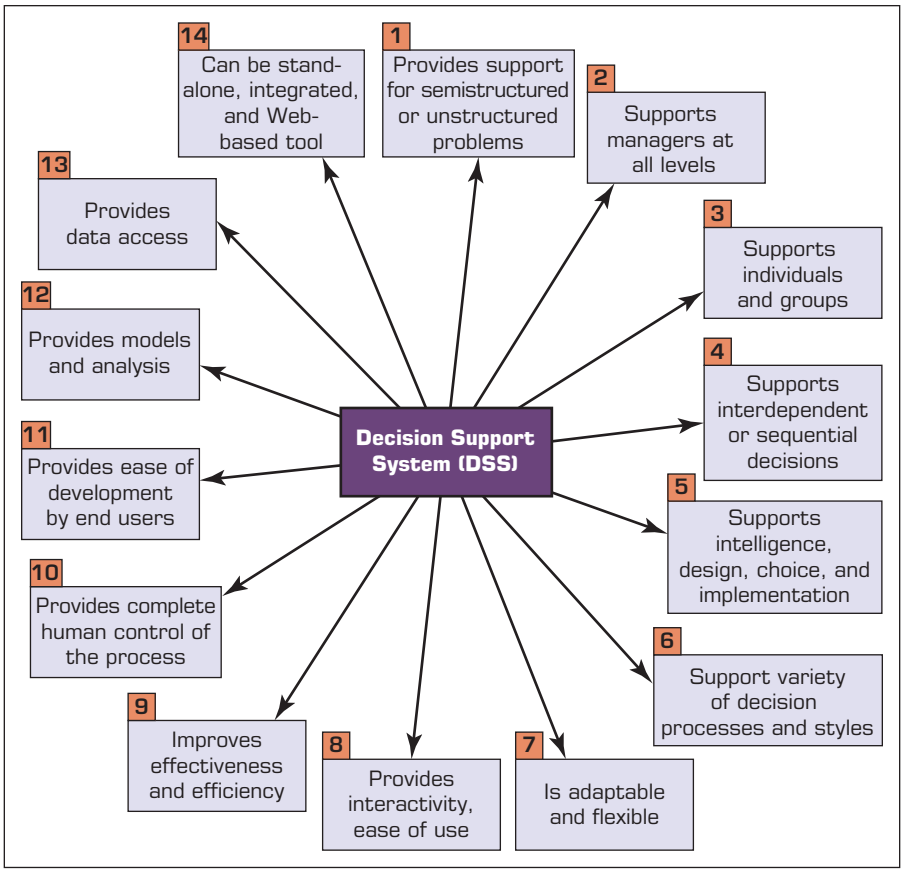


FIGURE 1.3 **Key Characteristics and Capabilities of DSS.**

end users) through an interactive and iterative process. It can support all phases of decision making and may include a knowledge component. Finally, a DSS can be used by a single user or can be Web based for use by many people at several locations.

THE CHARACTERISTICS AND CAPABILITIES OF DSS Because there is no consensus on exactly what a DSS is, there is obviously no agreement on the standard characteristics and capabilities of DSS. The capabilities in Figure 1.3 constitute an ideal set, some members of which are described in the definitions of DSS and illustrated in the application cases.

The key characteristics and capabilities of DSS (as shown in Figure 1.3) are as follows:

1. Supports decision makers, mainly in semistructured and unstructured situations, by bringing together human judgment and computerized information. Such problems cannot be solved (or cannot be solved conveniently) by other computerized systems or through use of standard quantitative methods or tools. Generally, these problems gain structure as the DSS is developed. Even some structured problems have been solved by DSS.
2. Supports all managerial levels, ranging from top executives to line managers.
3. Supports individuals as well as groups. Less-structured problems often require the involvement of individuals from different departments and organizational levels or even from different organizations. DSS supports virtual teams through collaborative Web tools. DSS has been developed to support individual and group work as well

as to support individual decision making and groups of decision makers working somewhat independently.

4. Supports interdependent and/or sequential decisions. The decisions may be made once, several times, or repeatedly.
5. Supports all phases of the decision-making process: intelligence, design, choice, and implementation.
6. Supports a variety of decision-making processes and styles.
7. Is flexible, so users can add, delete, combine, change, or rearrange basic elements. The decision maker should be reactive, able to confront changing conditions quickly, and able to adapt the DSS to meet these changes. It is also flexible in that it can be readily modified to solve other, similar problems.
8. Is user-friendly, has strong graphical capabilities, and a natural language interactive human-machine interface can greatly increase the effectiveness of DSS. Most new DSS applications use Web-based interfaces or mobile platform interfaces.
9. Improves the effectiveness of decision making (e.g., accuracy, timeliness, quality) rather than its efficiency (e.g., the cost of making decisions). When DSS is deployed, decision making often takes longer, but the decisions are better.
10. Provides complete control by the decision maker over all steps of the decision-making process in solving a problem. A DSS specifically aims to support, not to replace, the decision maker.
11. Enables end users to develop and modify simple systems by themselves. Larger systems can be built with assistance from IS specialists. Spreadsheet packages have been utilized in developing simpler systems. OLAP and data mining software in conjunction with data warehouses enable users to build fairly large, complex DSS.
12. Provides models that are generally utilized to analyze decision-making situations. The modeling capability enables experimentation with different strategies under different configurations.
13. Provides access to a variety of data sources, formats, and types, including GIS, multimedia, and object-oriented data.
14. Can be employed as a stand-alone tool used by an individual decision maker in one location or distributed throughout an organization and in several organizations along the supply chain. It can be integrated with other DSS and/or applications, and it can be distributed internally and externally, using networking and Web technologies.

These key DSS characteristics and capabilities allow decision makers to make better, more consistent decisions in a timely manner, and they are provided by major DSS components,

Components of a Decision Support System

A DSS application can be composed of a data management subsystem, a model management subsystem, a user interface subsystem, and a knowledge-based management subsystem. We show these in Figure 1.4.

The Data Management Subsystem

The data management subsystem includes a database that contains relevant data for the situation and is managed by software called the database management system (DBMS). *DBMS* is used as both singular and plural (*system* and *systems*) terms, as are many other acronyms in this text. The data management subsystem can be interconnected with the corporate data warehouse, a repository for corporate relevant decision-making data.

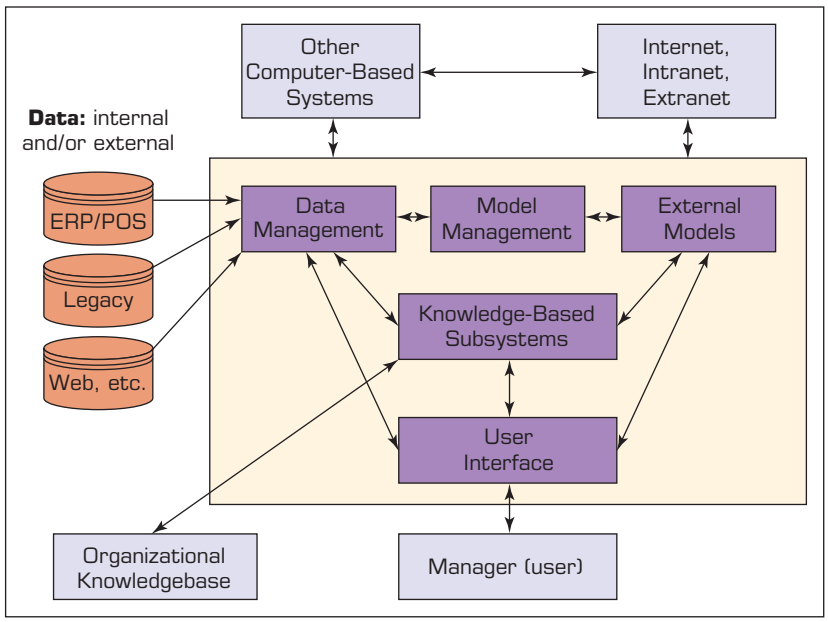


FIGURE 1.4 **Schematic View of DSS.**

Usually, the data are stored or accessed via a database Web server. The data management subsystem is composed of the following elements:

- DSS database
- Database management system
- Data directory
- Query facility

Many of the BI or descriptive analytics applications derive their strength from the data management side of the subsystems.

The Model Management Subsystem

The model management subsystem is the component that includes financial, statistical, management science, or other quantitative models that provide the system's analytical capabilities and appropriate software management. Modeling languages for building custom models are also included. This software is often called a model base management system (MBMS). This component can be connected to corporate or external storage of models. Model solution methods and management systems are implemented in Web development systems (such as Java) to run on application servers. The model management subsystem of a DSS is composed of the following elements:

- Model base
- MBMS
- Modeling language
- Model directory
- Model execution, integration, and command processor

Because DSS deals with semistructured or unstructured problems, it is often necessary to customize models, using programming tools and languages. Some examples of these are .NET Framework languages, C++, and Java. OLAP software may also be used to work with models in data analysis. Even languages for simulations such as Arena and

Application Case 1.2 SNAP DSS Helps OneNet Make Telecommunications Rate Decisions

Telecommunications network services to educational institutions and government entities are typically provided by a mix of private and public organizations. Many states in the United States have one or more state agencies that are responsible for providing network services to schools, colleges, and other state agencies. One example of such an agency is OneNet in Oklahoma. OneNet is a division of the Oklahoma State Regents for Higher Education and operated in cooperation with the Office of State Finance.

Usually agencies such as OneNet operate as an enterprise-type fund. They must recover their costs through billing their clients and/or by justifying appropriations directly from the state legislatures. This cost recovery should occur through a pricing mechanism that is efficient, simple to implement, and equitable. This pricing model typically needs to recognize many factors: convergence of voice, data, and video traffic on the same infrastructure; diversity of user base in terms of educational institutions and state agencies; diversity of applications in use by state clients from e-mail to videoconferences, IP telephoning, and distance learning; recovery of current costs as well as planning for upgrades and future developments; and leverage of the shared infrastructure to enable further economic development and collaborative work across the state that leads to innovative uses of OneNet.

These considerations led to the development of a spreadsheet-based model. The system, SNAP-DSS, or Service Network Application and Pricing (SNAP)-based DSS, was developed in Microsoft Excel 2007 and used the VBA programming language.

The SNAP-DSS offers OneNet the ability to select the rate card options that best fit the preferred pricing strategies by providing a real-time, user-friendly, graphical user interface (GUI). In addition, the SNAP-DSS not only illustrates the influence of the changes in the pricing factors on each rate card option but also allows the user to analyze various rate card options in different scenarios using different parameters. This model has been used by OneNet financial planners to gain insights into their customers and analyze many what-if scenarios of different rate plan options.

Source: Based on J. Chongwatpol and R. Sharda. (2010, December). "SNAP: A DSS to Analyze Network Service Pricing for State Networks." *Decision Support Systems, 50*(1), pp. 347–359.

statistical packages such as those of SPSS offer modeling tools developed through the use of a proprietary programming language. For small- and medium-sized DSS or for less complex ones, a spreadsheet (e.g., Excel) is usually used. We use Excel for several examples in this book. Application Case 1.2 describes a spreadsheet-based DSS.

The User Interface Subsystem

The user communicates with and commands the DSS through the user interface subsystem. The user is considered part of the system. Researchers assert that some of the unique contributions of DSS are derived from the intensive interaction between the computer and the decision maker. A difficult user interface is one of the major reasons that managers do not use computers and quantitative analyses as much as they could, given the availability of these technologies. The Web browser provided a familiar, consistent GUI structure for many DSS in the 2000s. For locally used DSS, a spreadsheet also provides a familiar user interface. The Web browser has been recognized as an effective DSS GUI because it is flexible, user-friendly, and a gateway to almost all sources of necessary information and data. Essentially, Web browsers have led to the development of portals and dashboards, which front end many DSS.

Explosive growth in portable devices, including smartphones and tablets, has changed the DSS user interfaces as well. These devices allow either handwritten input or

typed input from internal or external keyboards. Some DSS user interfaces utilize natural language input (i.e., text in a human language) so that the users can easily express themselves in a meaningful way. Cell phone inputs through short message service (SMS) or chatbots are becoming more common for at least some consumer DSS-type applications. For example, one can send an SMS request for search on any topic to GOOGL (46645). Such capabilities are most useful in locating nearby businesses, addresses, or phone numbers, but it can also be used for many other decision support tasks. For example, users can find definitions of words by entering the word "define" followed by a word, such as "define extenuate." Some of the other capabilities include

- Price lookups: "Price 64GB iPhone X."
- Currency conversions: "10 US dollars in euros."
- Sports scores and game times: Just enter the name of a team ("NYC Giants"), and Google SMS will send the most recent game's score and the date and time of the next match.

This type of SMS-based search capability is also available for other search engines such as Microsoft's search engine Bing.

With the emergence of smartphones such as Apple's iPhone and Android smartphones from many vendors, many companies are developing *apps* to provide purchasing-decision support. For example, Amazon's app allows a user to take a picture of any item in a store (or wherever) and send it to **Amazon.com**. **Amazon.com's** graphics-understanding algorithm tries to match the image to a real product in its databases and sends the user a page similar to **Amazon.com's** product info pages, allowing users to perform price comparisons in real time. Millions of other apps have been developed that provide consumers support for decision making on finding and selecting stores/restaurants/service providers on the basis of location, recommendations from others, and especially from your own social circles. Search activities noted in the previous paragraph are also largely accomplished now through apps provided by each search provider.

Voice input for these devices and the new smart speakers such as Amazon Echo (Alexa) and Google Home is common and fairly accurate (but not perfect). When voice input with accompanying speech-recognition software (and readily available text-to-speech software) is used, verbal instructions with accompanied actions and outputs can be invoked. These are readily available for DSS and are incorporated into the portable devices described earlier. An example of voice inputs that can be used for a general-purpose DSS is Apple's Siri application and Google's Google Now service. For example, a user can give her or his zip code and say "pizza delivery." These devices provide the search results and can even place a call to a business.

The Knowledge-Based Management Subsystem

Many of the user interface developments are closely tied to the major new advances in their knowledge-based systems. The knowledge-based management subsystem can support any of the other subsystems or act as an independent component. It provides intelligence to augment the decision maker's own or to help understand a user's query so as to provide a consistent answer. It can be interconnected with the organization's knowledge repository (part of a KMS), which is sometimes called the *organizational knowledge base*, or connect to thousands of external knowledge sources. Many artificial intelligence methods have been implemented in the current generation of learning systems and are easy to integrate into the other DSS components. One of the most widely publicized knowledge-based DSS is IBM's Watson, which was introduced in the opening vignette and will be described in more detail later.

This section has covered the history and progression of Decision Support Systems in brief. In the next section we discuss evolution of this support to business intelligence, analytics, and data science.

SECTION 1.3 REVIEW QUESTIONS

1. List and briefly describe Simon's four phases of decision making.
2. What is the difference between a problem and its symptoms?
3. Why is it important to classify a problem?
4. Define *implementation*.
5. What are structured, unstructured, and semistructured decisions? Provide two examples of each.
6. Define *operational control, managerial control*, and *strategic planning*. Provide two examples of each.
7. What are the nine cells of the decision framework? Explain what each is for.
8. How can computers provide support for making structured decisions?
9. How can computers provide support for making semistructured and unstructured decisions?

1.4 EVOLUTION OF COMPUTERIZED DECISION SUPPORT TO BUSINESS INTELLIGENCE/ANALYTICS/DATA SCIENCE

The timeline in Figure 1.5 shows the terminology used to describe analytics since the 1970s. During the 1970s, the primary focus of information systems support for decision making focused on providing structured, periodic reports that a manager could use for decision making (or ignore them). Businesses began to create routine reports to inform decision makers (managers) about what had happened in the previous period (e.g., day, week, month, quarter). Although it was useful to know what had happened in the past, managers needed more than this: They needed a variety of reports at different levels of granularity to better understand and address changing needs and challenges of the business. These were usually called *management information systems (MIS)*. In the early 1970s, Scott-Morton first articulated the major concepts of DSS. He defined DSS as "interactive computer-based systems, which help decision makers utilize *data* and *models* to solve unstructured problems" (Gorry and Scott-Morton, 1971). The following is another classic DSS definition provided by Keen and Scott-Morton (1978):

> Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems.

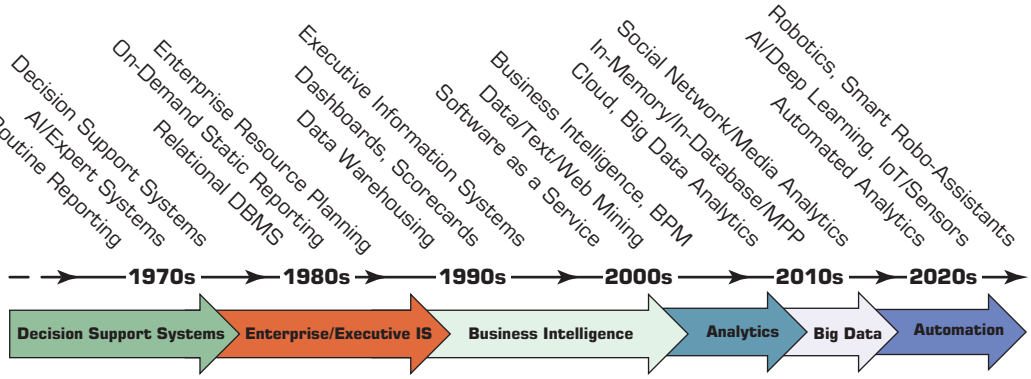


FIGURE 1.5 Evolution of Decision Support, Business Intelligence, Analytics, and AI.

Note that the term *decision support system*, like *management information system* and several other terms in the field of IT, is a content-free expression (i.e., it means different things to different people). Therefore, there is no universally accepted definition of DSS.

During the early days of analytics, data were often obtained from the domain experts using manual processes (i.e., interviews and surveys) to build mathematical or knowledge-based models to solve constrained optimization problems. The idea was to do the best with limited resources. Such decision support models were typically called operations research (OR). The problems that were too complex to solve optimally (using linear or nonlinear mathematical programming techniques) were tackled using heuristic methods such as simulation models. (We will introduce these as prescriptive analytics later in this chapter).

In the late 1970s and early 1980s, in addition to the mature OR models that were being used in many industries and government systems, a new and exciting line of models had emerged: rule-based expert systems (ESs). These systems promised to capture experts' knowledge in a format that computers could process (via a collection of if-then-else rules or heuristics) so that these could be used for consultation much the same way that one would use domain experts to identify a structured problem and to prescribe the most probable solution. ESs allowed scarce expertise to be made available where and when needed, using an "intelligent" DSS.

The 1980s saw a significant change in the way organizations captured business-related data. The old practice had been to have multiple disjointed information systems tailored to capture transactional data of different organizational units or functions (e.g., accounting, marketing and sales, finance, manufacturing). In the 1980s, these systems were integrated as enterprise-level information systems that we now commonly call *enterprise resource planning (ERP)* systems. The old mostly sequential and nonstandardized data representation schemas were replaced by relational database management (RDBM) systems. These systems made it possible to improve the capture and storage of data as well as the relationships between organizational data fields while significantly reducing the replication of information. The need for RDBM and ERP systems emerged when data integrity and consistency became an issue, significantly hindering the effectiveness of business practices. With ERP, all the data from every corner of the enterprise is collected and integrated into a consistent schema so that every part of the organization has access to the single version of the truth when and where needed. In addition to the emergence of ERP systems, or perhaps because of these systems, business reporting became an on-demand, as-needed business practice. Decision makers could decide when they needed to or wanted to create specialized reports to investigate organizational problems and opportunities.

In the 1990s, the need for more versatile reporting led to the development of executive information systems (EISs; DSS designed and developed specifically for executives and their decision-making needs). These systems were designed as graphical dashboards and scorecards so that they could serve as visually appealing displays while focusing on the most important factors for decision makers to keep track of the key performance indicators. To make this highly versatile reporting possible while keeping the transactional integrity of the business information systems intact, it was necessary to create a middle data tier known as a DW as a repository to specifically support business reporting and decision making. In a very short time, most large- to medium-sized businesses adopted data warehousing as their platform for enterprise-wide decision making. The dashboards and scorecards got their data from a DW, and by doing so, they were not hindering the efficiency of the business transaction systems mostly referred to as ERP systems.

In the 2000s, the DW-driven DSS began to be called *BI systems*. As the amount of longitudinal data accumulated in the DWs increased, so did the capabilities of hardware

and software to keep up with the rapidly changing and evolving needs of the decision makers. Because of the globalized competitive marketplace, decision makers needed current information in a very digestible format to address business problems and to take advantage of market opportunities in a timely manner. Because the data in a DW are updated periodically, they do not reflect the latest information. To elevate this information latency problem, DW vendors developed a system to update the data more frequently, which led to the terms *real-time data warehousing* and, more realistically, *right-time data warehousing*, which differs from the former by adopting a data-refreshing policy based on the needed freshness of the data items (i.e., not all data items need to be refreshed in real time). DWs are very large and feature rich, and it became necessary to "mine" the corporate data to "discover" new and useful knowledge nuggets to improve business processes and practices, hence, the terms *data mining* and *text mining*. With the increasing volumes and varieties of data, the needs for more storage and more processing power emerged. Although large corporations had the means to tackle this problem, small- to medium-sized companies needed more financially manageable business models. This need led to service-oriented architecture and software and infrastructure-as-a-service analytics business models. Smaller companies, therefore, gained access to analytics capabilities on an as-needed basis and paid only for what they used, as opposed to investing in financially prohibitive hardware and software resources.

In the 2010s, we are seeing yet another paradigm shift in the way that data are captured and used. Largely because of the widespread use of the Internet, new data generation mediums have emerged. Of all the new data sources (e.g., radio-frequency identification [RFID] tags, digital energy meters, clickstream Web logs, smart home devices, wearable health monitoring equipment), perhaps the most interesting and challenging is social networking/social media. These unstructured data are rich in information content, but analysis of such data sources poses significant challenges to computational systems from both software and hardware perspectives. Recently, the term *Big Data* has been coined to highlight the challenges that these new data streams have brought on us. Many advancements in both hardware (e.g., massively parallel processing with very large computational memory and highly parallel multiprocessor computing systems) and software/algorithms (e.g., Hadoop with MapReduce and NoSQL, Spark) have been developed to address the challenges of Big Data.

The last few years and the upcoming decade are bringing massive growth in many exciting dimensions. For example, streaming analytics and the sensor technologies have enabled the IoT. Artificial Intelligence is changing the shape of BI by enabling new ways of analyzing images through deep learning, not just traditional visualization of data. Deep learning and AI are also helping grow voice recognition and speech synthesis, leading to new interfaces in interacting with technologies. Almost half of U.S. households already have a smart speaker such as Amazon Echo or Google Home and have begun to interact with data and systems using voice interfaces. Growth in video interfaces will eventually enable gesture-based interaction with systems. All of these are being enabled due to massive cloud-based data storage and amazingly fast processing capabilities. And more is yet to come.

It is hard to predict what the next decade will bring and what the new analytics-related terms will be. The time between new paradigm shifts in information systems and particularly in analytics has been shrinking, and this trend will continue for the foreseeable future. Even though analytics is not new, the explosion in its popularity is very new. Thanks to the recent explosion in Big Data, ways to collect and store these data and intuitive software tools, data-driven insights are more accessible to business professionals than ever before. Therefore, in the midst of global competition, there is a huge opportunity to make better managerial decisions by using data and analytics to increase revenue while decreasing costs by building better products, improving customer experience, and catching fraud before it happens, improving customer engagement through targeting and customization, and developing entirely

new lines of business, all with the power of analytics and data. More and more companies are now preparing their employees with the know-how of business analytics to drive effectiveness and efficiency in their day-to-day decision-making processes.

The next section focuses on a framework for BI. Although most people would agree that BI has evolved into analytics and data science, many vendors and researchers still use that term. So the next few paragraphs pay homage to that history by specifically focusing on what has been called BI. Following the next section, we introduce analytics and use that as the label for classifying all related concepts.

A Framework for Business Intelligence

The decision support concepts presented in Sections 1.2 and 1.3 have been implemented incrementally, under different names, by many vendors that have created tools and methodologies for decision support. As noted in Section 1.2, as the enterprise-wide systems grew, managers were able to access user-friendly reports that enabled them to make decisions quickly. These systems, which were generally called EISs, then began to offer additional visualization, alerts, and performance measurement capabilities. By 2006, the major *commercial* products and services appeared under the term *business intelligence (BI)*.

DEFINITIONS OF BI *Business intelligence (BI)* is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies. It is, like DSS, a content-free expression, so it means different things to different people. Part of the confusion about BI lies in the flurry of acronyms and buzzwords that are associated with it (e.g., business performance management [BPM]). BI's major objective is to enable interactive access (sometimes in real time) to data, to enable manipulation of data, and to give business managers and analysts the ability to conduct appropriate analyses. By analyzing historical and current data, situations, and performances, decision makers get valuable insights that enable them to make more informed and better decisions. The process of BI is based on the *transformation* of data to information, then to decisions, and finally to actions.

A BRIEF HISTORY OF BI The term *BI* was coined by the Gartner Group in the mid-1990s. However, as the history in the previous section points out, the concept is much older; it has its roots in the MIS reporting systems of the 1970s. During that period, reporting systems were static, were two dimensional, and had no analytical capabilities. In the early 1980s, the concept of EISs emerged. This concept expanded the computerized support to top-level managers and executives. Some of the capabilities introduced were dynamic multidimensional (ad hoc or on-demand) reporting, forecasting and prediction, trend analysis, drill-down to details, status access, and critical success factors. These features appeared in dozens of commercial products until the mid-1990s. Then the same capabilities and some new ones appeared under the name BI. Today, a good BI-based enterprise information system contains all the information that executives need. So, the original concept of EIS was transformed into BI. By 2005, BI systems started to include *artificial intelligence* capabilities as well as powerful analytical capabilities. Figure 1.6 illustrates the various tools and techniques that may be included in a BI system. It illustrates the evolution of BI as well. The tools shown in Figure 1.6 provide the capabilities of BI. The most sophisticated BI products include most of these capabilities; others specialize in only some of them.

The Architecture of BI

A BI system has four major components: a *DW*, with its source data; *business analytics*, a collection of tools for manipulating, mining, and analyzing the data in the DW; *BPM* for monitoring and analyzing performance; and a *user interface* (e.g., a **dashboard**). The relationship among these components is illustrated in Figure 1.7.

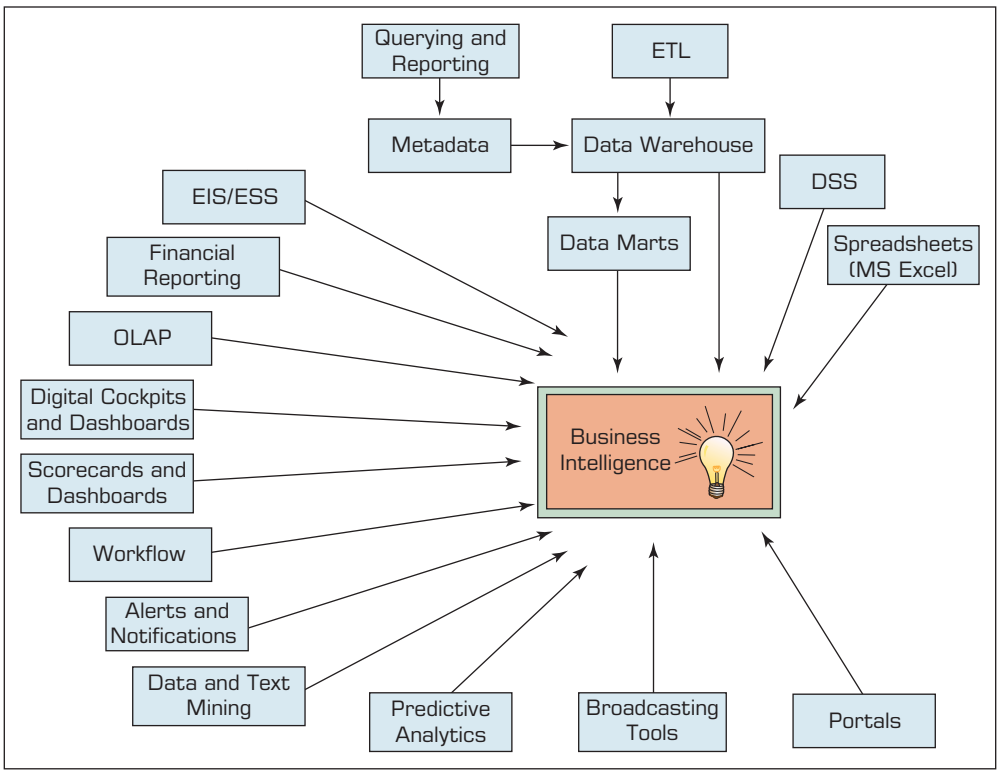


FIGURE 1.6 **Evolution of Business Intelligence (BI).**

The Origins and Drivers of BI

Where did modern approaches to DW and BI come from? What are their roots, and how do those roots affect the way organizations are managing these initiatives today? Today's investments in information technology are under increased scrutiny in terms of their bottom-line impact and potential. The same is true of DW and the BI applications that make these initiatives possible.

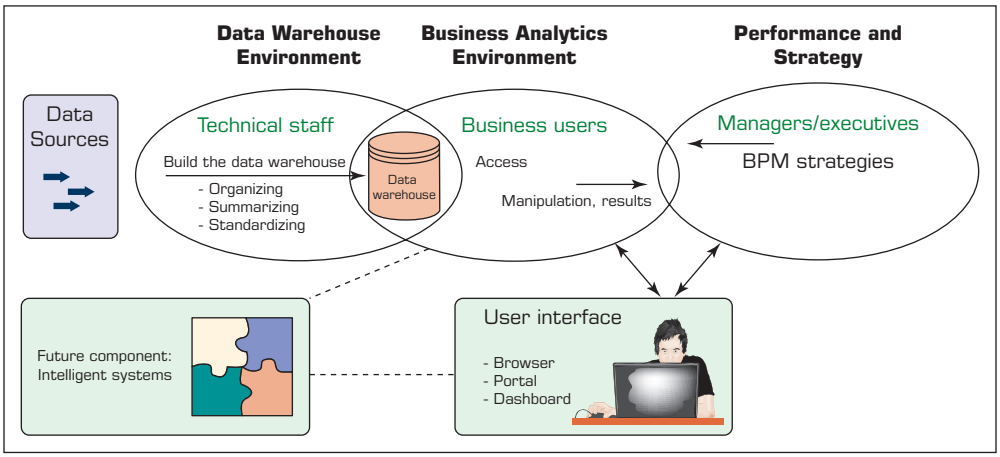


FIGURE 1.7 **A High-Level Architecture of BI.** *Source:* Based on W. Eckerson. (2003). *Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions.* Seattle, WA: The Data Warehousing Institute, p. 32, Illustration 5.

Organizations are being compelled to capture, understand, and harness their data to support decision making to improve business operations. Legislation and regulation (e.g., the Sarbanes-Oxley Act of 2002) now require business leaders to document their business processes and to sign off on the legitimacy of the information they rely on and report to stakeholders. Moreover, business cycle times are now extremely compressed; faster, more informed, and better decision making is, therefore, a competitive imperative. Managers need the *right information* at the *right time* and in the *right place*. This is the mantra for modern approaches to BI.

Organizations have to work smart. Paying careful attention to the management of BI initiatives is a necessary aspect of doing business. It is no surprise, then, that organizations are increasingly championing BI and under its new incarnation as analytics.

Data Warehouse as a Foundation for Business Intelligence

BI systems rely on a DW as the information source for creating insight and supporting managerial decisions. A multitude of organizational and external data is captured, transformed, and stored in a DW to support timely and accurate decisions through enriched business insight. In simple terms, a *DW* is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to managers throughout the organization. Data are usually structured to be available in a form ready for analytical processing activities (i.e., OLAP, data mining, querying, reporting, and other decision support applications). A DW is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process.

Whereas a DW is a repository of data, data warehousing is literally the entire process. Data warehousing is a discipline that results in applications that provide decision support capability, allows ready access to business information, and creates business insight. The three main types of data warehouses are data marts (DMs), operational data stores (ODS), and enterprise data warehouses (EDW). Whereas a DW combines databases across an entire enterprise, a DM is usually smaller and focuses on a particular subject or department. A DM is a subset of a data warehouse, typically consisting of a single subject area (e.g., marketing, operations). An operational data store (ODS) provides a fairly recent form of customer information file. This type of database is often used as an interim staging area for a DW. Unlike the static contents of a DW, the contents of an ODS are updated throughout the course of business operations. An EDW is a large-scale data warehouse that is used across the enterprise for decision support. The large-scale nature of an EDW provides integration of data from many sources into a standard format for effective BI and decision support applications. EDWs are used to provide data for many types of DSS, including CRM, supply chain management (SCM), BPM, business activity monitoring, product lifecycle management, revenue management, and sometimes even KMS.

In Figure 1.8, we show the DW concept. Data from many different sources can be extracted, transformed, and loaded into a DW for further access and analytics for decision support. Further details of DW are available in an online chapter on the book's Web site.

Transaction Processing versus Analytic Processing

To illustrate the major characteristics of BI, first we will show what BI is not—namely, transaction processing. We are all familiar with the information systems that support our transactions, like ATM withdrawals, bank deposits, and cash register scans at the grocery store. These *transaction processing* systems are constantly involved in handling updates to what we might call *operational databases*. For example, in an ATM withdrawal transaction, we need to reduce our bank balance accordingly; a bank deposit adds to an account; and a grocery store purchase is likely reflected in the store's calculation of total sales for the day, and it should reflect an appropriate reduction in the store's inventory for the items we bought, and so on. These **online transaction processing (OLTP)** systems handle a

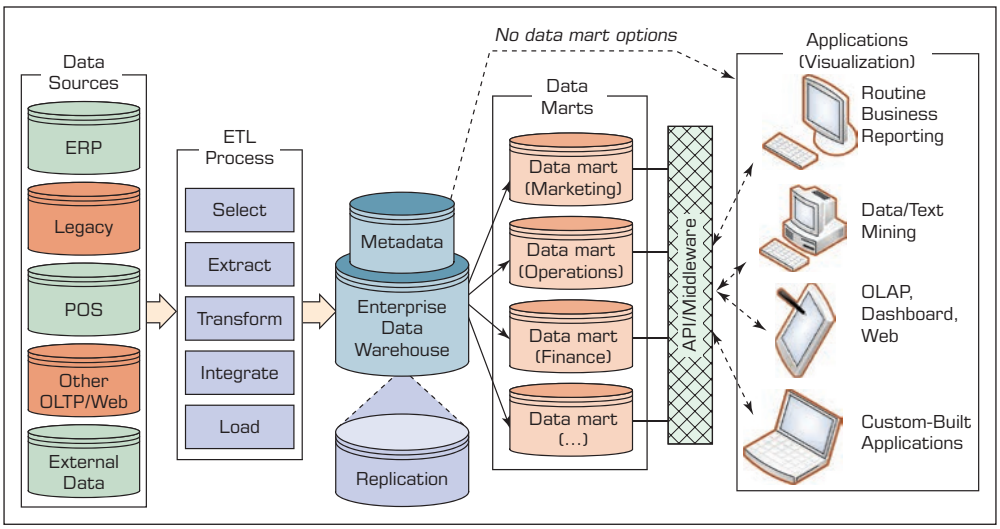


FIGURE 1.8 **Data Warehouse Framework and Views.**

company's routine ongoing business. In contrast, a DW is typically a distinct system that provides storage for data that will be used for *analysis*. The intent of that analysis is to give management the ability to scour data for information about the business, and it can be used to provide tactical or operational decision support whereby, for example, line personnel can make quicker and/or more informed decisions. DWs are intended to work with informational data used for **online analytical processing (OLAP)** systems.

Most operational data in ERP systems—and in their complementary siblings like *SCM* or *CRM*—are stored in an OLTP system, which is a type of computer processing where the computer responds immediately to user requests. Each request is considered to be a *transaction*, which is a computerized record of a discrete event, such as the receipt of inventory or a customer order. In other words, a transaction requires a set of two or more database updates that must be completed in an all-or-nothing fashion.

The very design that makes an OLTP system efficient for transaction processing makes it inefficient for end-user ad hoc reports, queries, and analysis. In the 1980s, many business users referred to their mainframes as "black holes" because all the information went into them, but none ever came back. All requests for reports had to be programmed by the IT staff, whereas only "precanned" reports could be generated on a scheduled basis, and ad hoc real-time querying was virtually impossible. Although the client/server-based ERP systems of the 1990s were somewhat more report friendly, they have still been a far cry from a desired usability by regular, nontechnical end users for things such as operational reporting and interactive analysis. To resolve these issues, the notions of DW and BI were created.

DWs contain a wide variety of data that present a coherent picture of business conditions at a single point in time. The idea was to create a database infrastructure that was always online and contained all the information from the OLTP systems, including historical data, but reorganized and structured in such a way that it was fast and efficient for querying, analysis, and decision support. Separating the OLTP from analysis and decision support enables the benefits of BI that were described earlier.

A Multimedia Exercise in Business Intelligence

TUN includes videos (similar to the television show *CSI*) to illustrate concepts of analytics in different industries. These are called "BSI Videos (Business Scenario Investigations)." Not only are these entertaining, but they also provide the class with some questions for discussion. For starters, please go to **https://www.teradatauniversitynetwork.com/Library/Items/BSI-The-Case-of-the-Misconnecting-Passengers/** or **www.youtube.**

com/watch?v=NXEL5F4_aKA. Watch the video that appears on YouTube. Essentially, you have to assume the role of a customer service center professional. An incoming flight is running late, and several passengers are likely to miss their connecting flights. There are seats on one outgoing flight that can accommodate two of the four passengers. Which two passengers should be given priority? You are given information about customers' profiles and relationships with the airline. Your decisions might change as you learn more about those customers' profiles.

Watch the video, pause it as appropriate, and answer the questions on which passengers should be given priority. Then resume the video to get more information. After the video is complete, you can see the slides related to this video and how the analysis was prepared on a slide set at **www.slideshare.net/teradata/bsi-how-we-did-it-the-case-of-the-misconnecting-passengers**.

This multimedia excursion provides an example of how additional available information through an enterprise DW can assist in decision making.

Although some people equate DSS with BI, these systems are not, at present, the same. It is interesting to note that some people believe that DSS is a part of BI—one of its analytical tools. Others think that BI is a special case of DSS that deals mostly with reporting, communication, and collaboration (a form of data-oriented DSS). Another explanation (Watson, 2005) is that BI is a result of a continuous revolution, and as such, DSS is one of BI's original elements. Further, as noted in the next section onward, in many circles, BI has been subsumed by the new terms *analytics* or *data science*.

APPROPRIATE PLANNING AND ALIGNMENT WITH THE BUSINESS STRATEGY First and foremost, the fundamental reasons for investing in BI must be aligned with the company's business strategy. BI cannot simply be a technical exercise for the information systems department. It has to serve as a way to change the manner in which the company conducts business by improving its business processes and transforming decision-making processes to be more data driven. Many BI consultants and practitioners involved in successful BI initiatives advise that a framework for planning is a necessary precondition. One framework, proposed by Gartner, Inc. (2004), decomposed planning and execution into *business, organization, functionality*, and *infrastructure* components. At the business and organizational levels, strategic and operational objectives must be defined while considering the available organizational skills to achieve those objectives. Issues of organizational culture surrounding BI initiatives and building enthusiasm for those initiatives and procedures for the intra-organizational sharing of BI best practices must be considered by upper management—with plans in place to prepare the organization for change. One of the first steps in that process is to assess the IS organization, the skill sets of the potential classes of users, and whether the culture is amenable to change. From this assessment, and assuming there are justification and the need to move ahead, a company can prepare a detailed action plan. Another critical issue for BI implementation success is the integration of several BI projects (most enterprises use several BI projects) among themselves and with the other IT systems in the organization and its business partners.

Gartner and many other analytics consulting organizations promoted the concept of a BI competence center that would serve the following functions:

- A center can demonstrate how BI is clearly linked to strategy and execution of strategy.
- A center can serve to encourage interaction between the potential business user communities and the IS organization.
- A center can serve as a repository and disseminator of best BI practices between and among the different lines of business.
- Standards of excellence in BI practices can be advocated and encouraged throughout the company.
- The IS organization can learn a great deal through interaction with the user communities, such as knowledge about the variety of types of analytical tools that are needed.

- The business user community and IS organization can better understand why the DW platform must be flexible enough to provide for changing business requirements.
- The center can help important stakeholders like high-level executives see how BI can play an important role.

Over the last 10 years, the idea of a BI competence center has been abandoned because many advanced technologies covered in this book have reduced the need for a central group to organize many of these functions. Basic BI has now evolved to a point where much of it can be done in "self-service" mode by the end users. For example, many data visualizations are easily accomplished by end users using the latest visualization packages (Chapter 3 will introduce some of these). As noted by Duncan (2016), the BI team would now be more focused on producing curated data sets to enable self-service BI. Because analytics is now permeating across the whole organization, the BI competency center could evolve into an analytics community of excellence to promote best practices and ensure overall alignment of analytics initiatives with organizational strategy.

BI tools sometimes needed to be integrated among themselves, creating synergy. The need for integration pushed software vendors to continuously add capabilities to their products. Customers who buy an all-in-one software package deal with only one vendor and do not have to deal with system connectivity. But they may lose the advantage of creating systems composed from the "best-of-breed" components. This led to major chaos in the BI market space. Many of the software tools that rode the BI wave (e.g., Savvion, Vitria, Tibco, MicroStrategy, Hyperion) have either been acquired by other companies or have expanded their offerings to take advantage of six key trends that have emerged since the initial wave of surge in business intelligence:

- Big Data.
- Focus on customer experience as opposed to just operational efficiency.
- Mobile and even newer user interfaces—visual, voice, mobile.
- Predictive and prescriptive analytics, machine learning, artificial intelligence.
- Migration to cloud.
- Much greater focus on security and privacy protection.

This book covers many of these topics in significant detail by giving examples of how the technologies are evolving and being applied, and the managerial implications.

SECTION 1.4 REVIEW QUESTIONS

1. List three of the terms that have been predecessors of analytics.
2. What was the primary difference between the systems called MIS, DSS, and Executive Information Systems?
3. Did DSS evolve into BI or vice versa?
4. Define *BI*.
5. List and describe the major components of BI.
6. Define *OLTP*.
7. Define *OLAP*.
8. List some of the implementation topics addressed by Gartner's report.
9. List some other success factors of BI.

1.5 ANALYTICS OVERVIEW

The word *analytics* has largely replaced the previous individual components of computerized decision support technologies that have been available under various labels in the past. Indeed, many practitioners and academics now use the word *analytics* in place of BI. Although many authors and consultants have defined it slightly differently, one can

view **analytics** as the process of developing actionable decisions or recommendations for actions based on insights generated from historical data. According to the Institute for Operations Research and Management Science (INFORMS), analytics represents the combination of computer technology, management science techniques, and statistics to solve real problems. Of course, many other organizations have proposed their own interpretations and motivations for analytics. For example, SAS Institute Inc. proposed eight levels of analytics that begin with standardized reports from a computer system. These reports essentially provide a sense of what is happening with an organization. Additional technologies have enabled us to create more customized reports that can be generated on an ad hoc basis. The next extension of reporting takes us to OLAP-type queries that allow a user to dig deeper and determine specific sources of concern or opportunities. Technologies available today can also automatically issue alerts for a decision maker when performance warrants such alerts. At a consumer level, we see such alerts for weather or other issues. But similar alerts can also be generated in specific settings when sales fall above or below a certain level within a certain time period or when the inventory for a specific product is running low. All of these applications are made possible through analysis and queries of data being collected by an organization. The next level of analysis might entail statistical analysis to better understand patterns. These can then be taken a step further to develop forecasts or models for predicting how customers might respond to a specific marketing campaign or ongoing service/product offerings. When an organization has a good view of what is happening and what is likely to happen, it can also employ other techniques to make the best decisions under the circumstances.

This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by INFORMS in proposing three levels of analytics. These three levels are identified as descriptive, predictive, and prescriptive. Figure 1.9 presents a graphical view of these three levels of analytics. It suggests that these three are somewhat independent steps and one type of analytics

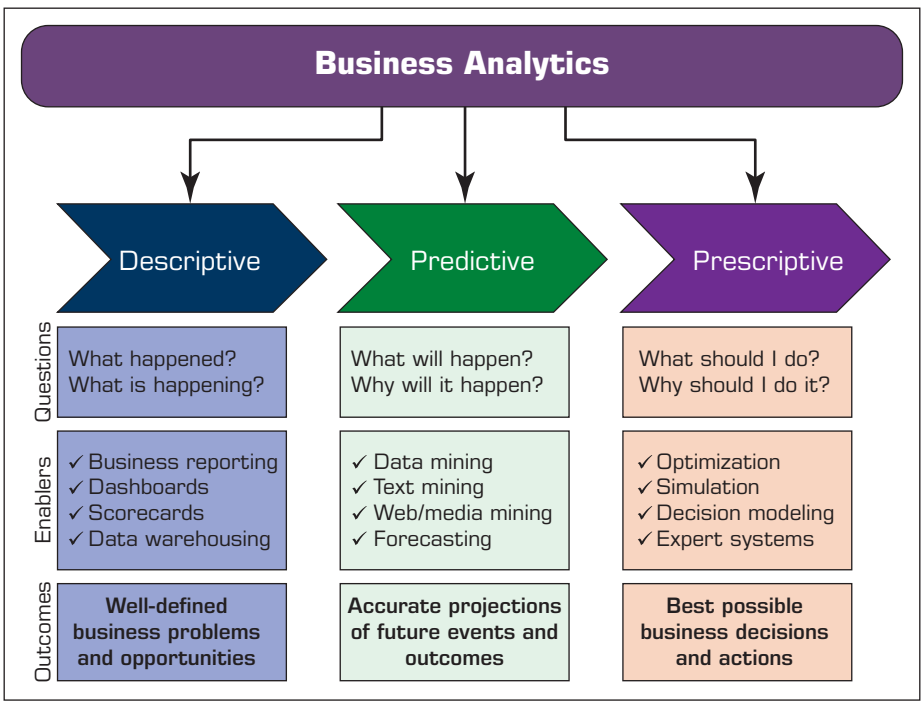


FIGURE 1.9 Three Types of Analytics.

applications leads to another. It also suggests that there is actually some overlap across these three types of analytics. In either case, the interconnected nature of different types of analytics applications is evident. We next introduce these three levels of analytics.

Descriptive Analytics

Descriptive (or reporting) analytics refers to knowing what is happening in the organization and understanding some underlying trends and causes of such occurrences. First, this involves the consolidation of data sources and availability of all relevant data in a form that enables appropriate reporting and analysis. Usually, the development of this data infrastructure is part of DWs. From this data infrastructure, we can develop appropriate reports, queries, alerts, and trends using various reporting tools and techniques.

A significant technology that has become a key player in this area is visualization. Using the latest visualization tools in the marketplace, we can now develop powerful insights in the operations of our organization. Application Cases 1.3 and 1.4 highlight some such applications.

Application Case 1.3

Silvaris Increases Business with Visual Analysis and Real-Time Reporting Capabilities

Silvaris Corporation was founded in 2000 by a team of forest industry professionals to provide technological advancement in the lumber and building material sector. Silvaris is the first e-commerce platform in the United States specifically for forest products and is headquartered in Seattle, Washington. It is a leading wholesale provider of industrial wood products and surplus building materials.

Silvaris sells its products and provides international logistics services to more than 3,500 customers. To manage various processes that are involved in a transaction, the company created a proprietary online trading platform to track information flow related to transactions between traders, accounting, credit, and logistics. This allowed Silvaris to share its real-time information with its customers and partners. But due to the rapidly changing prices of materials, it became necessary for Silvaris to get a real-time view of data without moving them into a separate reporting format.

Silvaris started using Tableau because of its ability to connect with and visualize live data. With dashboards created by Tableau that are easy to understand and explain, Silvaris started using it for reporting purposes. This helped Silvaris in pulling out information quickly from the data and identifying issues that impact its business. Silvaris succeeded in managing online versus offline orders with the help of reports generated by Tableau. Now, Silvaris keeps track of online orders placed by customers and knows when to send renew pushes to which customers to keep them purchasing online. Also, analysts of Silvaris can save time by generating dashboards instead of writing hundreds of pages of reports by using Tableau.

Sources: Tableau.com. "Silvaris Augments Proprietary Technology Platform with Tableau's Real-Time Reporting Capabilities." **http://www.tableau.com/sites/default/files/case-studies/silvaris-business-dashboards_0.pdf** (accessed September 2018); **Silvaris.com**. **http://www.silvaris.com** (accessed September 2018).

Questions for Case 1.3

1. What was the challenge faced by Silvaris?
2. How did Silvaris solve its problem using data visualization with Tableau?

What We Can Learn from This Application Case

Many industries need to analyze data in real time. Real-time analysis enables the analysts to identify issues that impact their business. Visualization is sometimes the best way to begin analyzing the live data streams. Tableau is one such data visualization tool that has the capability to analyze live data without bringing live data into a separate reporting format.

Application Case 1.4 Siemens Reduces Cost with the Use of Data Visualization

Siemens is a German company headquartered in Berlin, Germany. It is one of the world's largest companies focusing on the areas of electrification, automation, and digitalization. It has an annual revenue of 76 billion euros.

The visual analytics group of Siemens is tasked with end-to-end reporting solutions and consulting for all of Siemens internal BI needs. This group was facing the challenge of providing reporting solutions to the entire Siemens organization across different departments while maintaining a balance between governance and self-service capabilities. Siemens needed a platform that could analyze its multiple cases of customer satisfaction surveys, logistic processes, and financial reporting. This platform should be easy to use for their employees so that they could use these data for analysis and decision making. In addition, the platform should be easily integrated with existing Siemens systems and give employees a seamless user experience.

Siemens started using Dundas BI, a leading global provider of BI and data visualization solutions. It allowed Siemens to create highly interactive dashboards that enabled it to detect issues early and thus save a significant amount of money. The dashboards developed by Dundas BI helped Siemens global logistics organization answer questions like how different supply rates at different locations affect the operation, thus helping the company reduce cycle time by 12 percent and scrap cost by 25 percent.

Questions for Case 1.4

1. What challenges were faced by Siemens visual analytics group?
2. How did the data visualization tool Dundas BI help Siemens in reducing cost?

What We Can Learn from This Application Case

Many organizations want tools that can be used to analyze data from multiple divisions. These tools can help them improve performance and make data discovery transparent to their users so that they can identify issues within the business easily.

Sources: **Dundas.com**. "How Siemens Drastically Reduced Cost with Managed BI Applications." **https://www.dundas.com/Content/pdf/siemens-case-study.pdf** (accessed September 2018); Wikipedia.org. "SIEMENS." **https://en.wikipedia.org/wiki/Siemens** (accessed September 2018); **Siemens.com**. "About Siemens." **http://www.siemens.com/about/en/** (accessed September 2018).

Predictive Analytics

Predictive analytics aims to determine what is likely to happen in the future. This analysis is based on statistical techniques as well as other more recently developed techniques that fall under the general category of **data mining**. The goal of these techniques is to be able to predict whether the customer is likely to switch to a competitor ("churn"), what and how much the customer would likely buy next, what promotions the customer would respond to, whether the customer is a creditworthy risk, and so forth. A number of techniques are used in developing predictive analytical applications, including various classification algorithms. For example, as described in Chapters 4 and 5, we can use classification techniques such as logistic regression, decision tree models, and neural networks to predict how well a motion picture will do at the box office. We can also use clustering algorithms for segmenting customers into different clusters to be able to target specific promotions to them. Finally, we can use association mining techniques (Chapters 4 and 5) to estimate relationships between different purchasing behaviors. That is, if a customer buys one product, what else is the customer likely to purchase? Such analysis can assist a retailer in recommending or promoting related products. For example, any product search on **Amazon.com** results in the retailer also suggesting similar products that a customer may be interested in. We will study these techniques and their applications in Chapters 3 through 6. Application Case 1.5 illustrates one such application in sports.

Application Case 1.5 Analyzing Athletic Injuries

Any athletic activity is prone to injuries. If the injuries are not handled properly, then the team suffers. Using analytics to understand injuries can help in deriving valuable insights that would enable coaches and team doctors to manage the team composition, understand player profiles, and ultimately aid in better decision making concerning which players might be available to play at any given time.

In an exploratory study, Oklahoma State University analyzed U.S. football-related sports injuries by using reporting and predictive analytics. The project followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology (to be described in Chapter 4) to understand the problem of making recommendations on managing injuries, understanding the various data elements collected about injuries, cleaning the data, developing visualizations to draw various inferences, building predictive models to analyze the injury healing time period, and drawing sequence rules to predict the relationships among the injuries and the various body part parts afflicted with injuries.

The injury data set consisted of more than 560 football injury records, which were categorized into injury-specific variables—body part/site/laterality, action taken, severity, injury type, injury start and healing dates—and player/sport-specific variables—player ID, position played, activity, onset, and game location. Healing time was calculated for each record, which was classified into different sets of time periods: 0–1 month, 1–2 months, 2–4 months, 4–6 months, and 6–24 months.

Various visualizations were built to draw inferences from injury–data set information depicting the healing time period associated with players' positions, severity of injuries and the healing time period, treatment offered and the associated healing time period, major injuries afflicting body parts, and so forth.

Neural network models were built to predict each of the healing categories using IBM SPSS Modeler. Some of the predictor variables were current status of injury, severity, body part, body site, type of injury, activity, event location, action taken, and position played. The success of classifying the healing category was quite good: Accuracy was 79.6 percent. Based on the analysis, many recommendations were suggested, including employing more specialists' input from injury onset instead of letting the training room staff screen the injured players; training players at defensive positions to avoid being injured; and holding practice to thoroughly safety-check mechanisms.

Sources: "Sharda, R., Asamoah, D., & Ponna, N. (2013). "Research and Pedagogy in Business Analytics: Opportunities and Illustrative Examples." *Journal of Computing and Information Technology,* 21(3), pp. 171–182.

Questions for Case 1.5

1. What types of analytics are applied in the injury analysis?
2. How do visualizations aid in understanding the data and delivering insights into the data?
3. What is a classification problem?
4. What can be derived by performing sequence analysis?

What We Can Learn from This Application Case

For any analytics project, it is always important to understand the business domain and the current state of the business problem through extensive analysis of the only resource—historical data. Visualizations often provide a great tool for gaining the initial insights into data, which can be further refined based on expert opinions to identify the relative importance of the data elements related to the problem. Visualizations also aid in generating ideas for obscure problems, which can be pursued in building PMs that could help organizations in decision making.

Prescriptive Analytics

The third category of analytics is termed **prescriptive analytics**. The goal of prescriptive analytics is to recognize what is going on as well as the likely forecast and make decisions to achieve the best performance possible. This group of techniques has historically been studied under the umbrella of OR or management sciences and is generally aimed at

optimizing the performance of a system. The goal here is to provide a decision or a recommendation for a specific action. These recommendations can be in the form of a specific yes/no decision for a problem, a specific amount (say, price for a specific item or airfare to charge), or a complete set of production plans. The decisions may be presented to a decision maker in a report or may be used directly in an automated decision rules system (e.g., in airline pricing systems). Thus, these types of analytics can also be termed **decision or normative analytics**. Application Case 1.6 gives an example of such prescriptive analytic applications. We will learn about some aspects of prescriptive analytics in Chapter 8.

ANALYTICS APPLIED TO DIFFERENT DOMAINS Applications of analytics in various industry sectors have spawned many related areas or at least buzzwords. It is almost fashionable to attach the word *analytics* to any specific industry or type of data. Besides the general category of text analytics—aimed at getting value out of text (to be studied in Chapter 7)—or Web analytics—analyzing Web data streams (also in

Application Case 1.6

A Specialty Steel Bar Company Uses Analytics to Determine Available-to-Promise Dates

This application case is based on a project that involved one of the coauthors A company that does not wish to disclose its name (or even its precise industry) was facing a major problem of making decisions on which inventory of raw materials to use to satisfy which customers. This company supplies custom configured steel bars to its customers. These bars may be cut into specific shapes or sizes and may have unique material and finishing requirements. The company procures raw materials from around the world and stores them in its warehouse. When a prospective customer calls the company to request a quote for the specialty bars meeting specific material requirements (composition, origin of the metal, quality, shapes, sizes, etc.), the salesperson usually has just a little bit of time to submit such a quote including the date when the product can be delivered and, of course, prices, and so on. It must make available-to-promise (ATP) decisions, which determine in real time the dates when the salesperson can promise delivery of products that customers requested during the quotation stage. Previously, a salesperson had to make such decisions by analyzing reports on available inventory of raw materials. Some of the available raw material may have already been committed to another customer's order. Thus, the inventory in stock might not really be inventory available. On the other hand, there may be raw material that is expected to be delivered in the near future that could also be used for satisfying the order from this prospective customer. Finally, there might even be an opportunity to charge a premium for a new order by repurposing previously committed inventory to satisfy this new order while delaying an already committed order. Of course, such decisions should be based on the cost–benefit analyses of delaying a previous order. The system should thus be able to pull real-time data about inventory, committed orders, incoming raw material, production constraints, and so on.

To support these ATP decisions, a real-time DSS was developed to find an optimal assignment of the available inventory and to support additional what-if analysis. The DSS uses a suite of mixed-integer programming models that are solved using commercial software. The company has incorporated the DSS into its enterprise resource planning system to seamlessly facilitate its use of business analytics.

Questions for Case 1.6

1. Why would reallocation of inventory from one customer to another be a major issue for discussion?
2. How could a DSS help make these decisions?

Source: M. Pajouh Foad, D. Xing, S. Hariharan, Y. Zhou, B. Balasundaram, T. Liu, & R. Sharda, R. (2013). "Available-to-Promise in Practice: An Application of Analytics in the Specialty Steel Bar Products Industry." *Interfaces, 43*(6), pp. 503–517. **http://dx.doi.org/10.1287/inte.2013.0693** (accessed September 2018).

Chapter 7)—many industry- or problem-specific analytics professions/streams have been developed. Examples of such areas are marketing analytics, retail analytics, fraud analytics, transportation analytics, health analytics, sports analytics, talent analytics, behavioral analytics, and so forth. For example, we will soon see several applications in *sports analytics*. Application Case 1.5 could also be termed a case study in health analytics. The next section will introduce health analytics and market analytics broadly. Literally, any systematic analysis of data in a specific sector is being labeled as "(fill-in-blanks)" analytics. Although this may result in overselling the concept of analytics, the benefit is that more people in specific industries are aware of the power and potential of analytics. It also provides a focus to professionals developing and applying the concepts of analytics in a vertical sector. Although many of the techniques to develop analytics applications may be common, there are unique issues within each vertical segment that influence how the data may be collected, processed, analyzed, and the applications implemented. Thus, the differentiation of analytics based on a vertical focus is good for the overall growth of the discipline.

ANALYTICS OR DATA SCIENCE? Even as the concept of analytics is receiving more attention in industry and academic circles, another term has already been introduced and is becoming popular. The new term is *data science*. Thus, the practitioners of data science are data scientists. D. J. Patil of LinkedIn is sometimes credited with creating the term *data science*. There have been some attempts to describe the differences between data analysts and data scientists (e.g., see "Data Science Revealed," 2018) (**emc.com/collateral/about/news/emc-data-science-study-wp.pdf**). One view is that *data analyst* is just another term for professionals who were doing BI in the form of data compilation, cleaning, reporting, and perhaps some visualization. Their skill sets included Excel use, some SQL knowledge, and reporting. You would recognize those capabilities as descriptive or reporting analytics. In contrast, data scientists are responsible for predictive analysis, statistical analysis, and use of more advanced analytical tools and algorithms. They may have a deeper knowledge of algorithms and may recognize them under various labels—data mining, knowledge discovery, or machine learning. Some of these professionals may also need deeper programming knowledge to be able to write code for data cleaning/analysis in current Web-oriented languages such as Java or Python and statistical languages such as R. Many analytics professionals also need to build significant expertise in statistical modeling, experimentation, and analysis. Again, our readers should recognize that these fall under the predictive and prescriptive analytics umbrella. However, prescriptive analytics also includes more significant expertise in OR including optimization, simulation, and decision analysis. Those who cover these fields are more likely to be called *data scientists* than *analytics professionals*.

Our view is that the distinction between analytics professional and data scientist is more of a degree of technical knowledge and skill sets than functions. It may also be more of a distinction across disciplines. Computer science, statistics, and applied mathematics programs appear to prefer the data science label, reserving the analytics label for more business-oriented professionals. As another example of this, applied physics professionals have proposed using *network science* as the term for describing analytics that relate to groups of people—social networks, supply chain networks, and so forth. See **http://barabasi.com/networksciencebook/** for an evolving textbook on this topic.

Aside from a clear difference in the skill sets of professionals who only have to do descriptive/reporting analytics versus those who engage in all three types of analytics, the distinction between the two labels is fuzzy at best. We observe that graduates of our analytics programs tend to be responsible for tasks that are more in line with data

science professionals (as defined by some circles) than just reporting analytics. This book is clearly aimed at introducing the capabilities and functionality of all analytics (which include data science), not just reporting analytics. From now on, we will use these terms interchangeably.

WHAT IS BIG DATA? Any book on analytics and data science has to include significant coverage of what is called **Big Data analytics**. We cover it in Chapter 9 but here is a very brief introduction. Our brains work extremely quickly and efficiently and are versatile in processing large amounts of all kinds of data: images, text, sounds, smells, and video. We process all different forms of data relatively easily. Computers, on the other hand, are still finding it hard to keep up with the pace at which data are generated, let alone analyze them quickly. This is why we have the problem of Big Data. So, what is Big Data? Simply put, Big Data refers to data that cannot be stored in a single storage unit. Big Data typically refers to data that come in many different forms: structured, unstructured, in a stream, and so forth. Major sources of such data are clickstreams from Web sites, postings on social media sites such as Facebook, and data from traffic, sensors, or weather. A Web search engine such as Google needs to search and index billions of Web pages to give you relevant search results in a fraction of a second. Although this is not done in real time, generating an index of all the Web pages on the Internet is not an easy task. Luckily for Google, it was able to solve this problem. Among other tools, it has employed Big Data analytical techniques.

There are two aspects to managing data on this scale: storing and processing. If we could purchase an extremely expensive storage solution to store all this at one place on one unit, making this unit fault tolerant would involve a major expense. An ingenious solution was proposed that involved storing these data in chunks on different machines connected by a network—putting a copy or two of this chunk in different locations on the network, both logically and physically. It was originally used at Google (then called the Google File System) and later developed and released by an Apache project as the Hadoop Distributed File System (HDFS).

However, storing these data is only half of the problem. Data are worthless if they do not provide business value, and for them to provide business value, they must be analyzed. How can such vast amounts of data be analyzed? Passing all computation to one powerful computer does not work; this scale would create a huge overhead on such a powerful computer. Another ingenious solution was proposed: Push computation to the data instead of pushing data to a computing node. This was a new paradigm and gave rise to a whole new way of processing data. This is what we know today as the MapReduce programming paradigm, which made processing Big Data a reality. MapReduce was originally developed at Google, and a subsequent version was released by the Apache project called *Hadoop MapReduce*.

Today, when we talk about storing, processing, or analyzing Big Data, HDFS and MapReduce are involved at some level. Other relevant standards and software solutions have been proposed. Although the major toolkit is available as an open source, several companies have been launched to provide training or specialized analytical hardware or software services in this space. Some examples are HortonWorks, Cloudera, and Teradata Aster.

Over the past few years, what was called Big Data changed more and more as Big Data applications appeared. The need to process data coming in at a rapid rate added velocity to the equation. An example of fast data processing is algorithmic trading. This uses electronic platforms based on algorithms for trading shares on the financial market, which operates in microseconds. The need to process different kinds of data added variety to the equation. Another example of a wide variety of data is sentiment analysis, which

uses various forms of data from social media platforms and customer responses to gauge sentiments. Today, Big Data is associated with almost any kind of large data that have the characteristics of volume, velocity, and variety. As noted before, these are evolving quickly to encompass stream analytics, IoT, cloud computing, and deep learning–enabled AI. We will study these in various chapters in the book.

SECTION 1.5 REVIEW QUESTIONS

1. Define *analytics*.
2. What is descriptive analytics? What are the various tools that are employed in descriptive analytics?
3. How is descriptive analytics different from traditional reporting?
4. What is a DW? How can DW technology help enable analytics?
5. What is predictive analytics? How can organizations employ predictive analytics?
6. What is prescriptive analytics? What kinds of problems can be solved by prescriptive analytics?
7. Define *modeling* from the analytics perspective.
8. Is it a good idea to follow a hierarchy of descriptive and predictive analytics before applying prescriptive analytics?
9. How can analytics aid in objective decision making?
10. What is Big Data analytics?
11. What are the sources of Big Data?
12. What are the characteristics of Big Data?
13. What processing technique is applied to process Big Data?

1.6 ANALYTICS EXAMPLES IN SELECTED DOMAINS

You will see examples of analytics applications throughout various chapters. That is one of the primary approaches (exposure) of this book. In this section, we highlight three application areas—sports, healthcare, and retail—where there have been the most reported applications and successes.

Sports Analytics—An Exciting Frontier for Learning and Understanding Applications of Analytics

The application of analytics to business problems is a key skill, one that you will learn in this book. Many of these techniques are now being applied to improve decision making in all aspects of sports, a very hot area called *sports analytics*. It is the art and science of gathering data about athletes and teams to create insights that improve sports decisions, such as deciding which players to recruit, how much to pay them, who to play, how to train them, how to keep them healthy, and when they should be traded or retired. For teams, it involves business decisions such as ticket pricing as well as roster decisions, analysis of each competitor's strengths and weaknesses, and many game-day decisions.

Indeed, sports analytics is becoming a specialty within analytics. It is an important area because sport is a big business, generating about $145 billion in revenues each year plus an additional $100 billion in legal and $300 billion in illegal gambling, according to Price Waterhouse ("Changing the Game: Outlook for the Global Sports Market to 2015" (2015)). In 2014, only $125 million was spent on analytics (less than 0.1 percent

of revenues). This is expected to grow at a healthy rate to $4.7 billion by 2021 ("Sports Analytics Market Worth $4.7B by 2021" (2015)).

The use of analytics for sports was popularized by the *Moneyball* book by Michael Lewis in 2003 and the movie starring Brad Pitt in 2011. It showcased Oakland A's general manager Billy Beane and his use of data and analytics to turn a losing team into a winner. In particular, he hired an analyst who used analytics to draft players who were able to get on base as opposed to players who excelled at traditional measures like runs batted in or stolen bases. These insights allowed the team to draft prospects overlooked by other teams at reasonable starting salaries. It worked—the team made it to the playoffs in 2002 and 2003.

Now analytics are being used in all parts of sports. The analytics can be divided between the front office and back office. A good description with 30 examples appears in Tom Davenport's survey article (). Front-office business analytics include analyzing fan behavior ranging from predictive models for season ticket renewals and regular ticket sales to scoring tweets by fans regarding the team, athletes, coaches, and owners. This is very similar to traditional CRM. Financial analysis is also a key area such as when salary cap (for pros) or scholarship (for colleges) limits are part of the equation.

Back-office uses include analysis of both individual athletes and team play. For individual players, there is a focus on recruitment models and scouting analytics, analytics for strength and fitness as well as development, and PMs for avoiding overtraining and injuries. Concussion research is a hot field. Team analytics include strategies and tactics, competitive assessments, and optimal roster choices under various on-field or on-court situations.

The following representative examples illustrate how two sports organizations use data and analytics to improve sports operations in the same way that analytics have improved traditional industry decision making.

Example 1: The Business Office

Dave Ward works as a business analyst for a major pro baseball team, focusing on revenue. He analyzes ticket sales, both from season ticket holders and single-ticket buyers. Sample questions in his area of responsibility include why season ticket holders renew (or do not renew) their tickets as well as what factors drive last-minute individual seat ticket purchases. Another question is how to price the tickets.

Some of the analytical techniques Dave uses include simple statistics on fan behavior such as overall attendance and answers to survey questions about likelihood to purchase again. However, what fans say versus what they do can be different. Dave runs a survey of fans by ticket seat location ("tier") and asks about their likelihood of renewing their season tickets. But when he compares what they say versus what they do, he discovers big differences. (See Figure 1.10.) He found that 69 percent of fans in Tier 1 seats who said on the survey that they would "probably not renew" actually did. This

Tier	Highly Likely	Likely	Maybe	Probably Not	Certainly Not
1	92	88	75	69	45
2	88	81	70	65	38
3	80	76	68	55	36
4	77	72	65	45	25
5	75	70	60	35	25

FIGURE 1.10 Season Ticket Renewals—Survey Scores.

is useful insight that leads to action—customers in the green cells are the most likely to renew tickets and so require fewer marketing touches and dollars to convert compared to customers in the blue cells.

However, many factors influence fan ticket purchase behavior, especially price, which drives more sophisticated statistics and data analysis. For both areas, but especially single-game tickets, Dave is driving the use of dynamic pricing—moving the business from simple static pricing by seat location tier to day-by-day up-and-down pricing of individual seats. This is a rich research area for many sports teams and has huge upside potential for revenue enhancement. For example, his pricing takes into account the team's record, who they are playing, game dates and times, which star athletes play for each team, each fan's history of renewing season tickets or buying single tickets, and factors such as seat location, number of seats, and real-time information like traffic congestion historically at game time and even the weather. See Figure 1.11.

Which of these factors are important and by how much? Given his extensive statistics background, Dave builds regression models to pick out key factors driving these historic behaviors and create PMs to identify how to spend marketing resources to drive revenues. He builds churn models for season ticket holders to create segments of customers who will renew, will not renew, or are fence-sitters, which then drives more refined marketing campaigns.

In addition, Dave does sentiment scoring on fan comments such as tweets that help him segment fans into different loyalty segments. Other studies about single-game attendance drivers help the marketing department understand the impact of giveaways like bobble-heads or T-shirts or suggestions on where to make spot TV ad buys.

Beyond revenues, there are many other analytical areas that Dave's team works on, including merchandising, TV and radio broadcast revenues, inputs to the general manager on salary negotiations, draft analytics especially given salary caps, promotion effectiveness including advertising channels, and brand awareness, as well as partner analytics. He's a very busy guy!

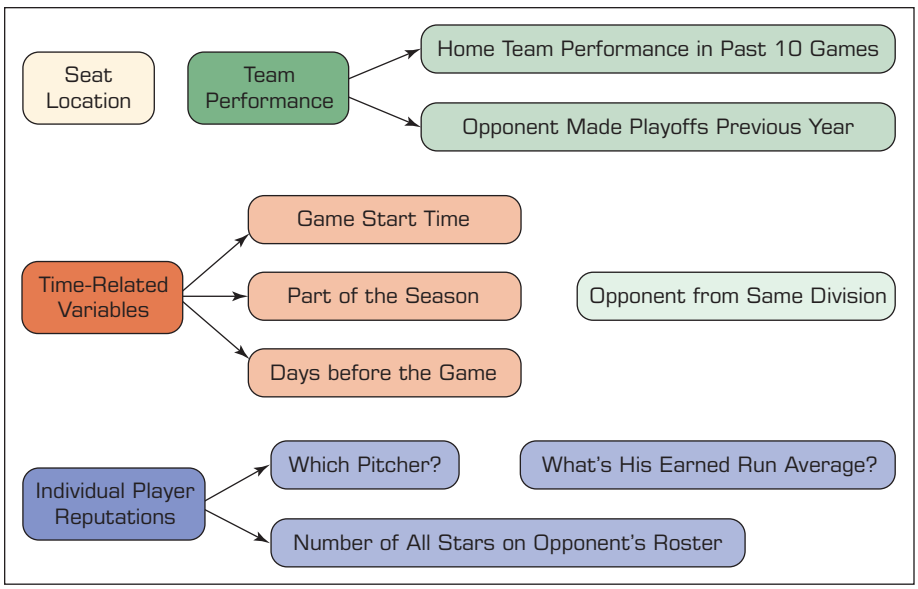


FIGURE 1.11 Dynamic Pricing Previous Work—Major League Baseball. *Source:* Based on C. Kemper and C. Breuer, "How Efficient is Dynamic Pricing for Sports Events? Designing a Dynamic Pricing Model for Bayern Munich", *Intl. Journal of Sports Finance*, 11, pp. 4–25, 2016.

Example 2: The Coach

Bob Breedlove is the football coach for a major college team. For him, everything is about winning games. His areas of focus include recruiting the best high school players, developing them to fit his offense and defense systems, and getting maximum effort from them on game days. Sample questions in his area of responsibility include: Whom do we recruit? What drills help develop their skills? How hard do I push our athletes? Where are opponents strong or weak, and how do we figure out their play tendencies?

Fortunately, his team has hired a new team operations expert, Dar Beranek, who specializes in helping the coaches make tactical decisions. She is working with a team of student interns who are creating opponent analytics. They used the coach's annotated game film to build a cascaded decision tree model (Figure 1.12) to predict whether the next play will be a running play or passing play. For the defensive coordinator, they have built heat maps (Figure 1.13) of each opponent's passing offense, illustrating their tendencies to throw left or right and into which defensive coverage zones. Finally, they built some time-series analytics (Figure 1.14) on explosive plays (defined as a gain of more than 16 yards for a passing play or more than 12 yards for a run play). For each play, they compare the outcome with their own defensive formations and the other team's offensive formations, which help Coach Breedlove react more quickly to formation shifts during a game. We explain the analytical techniques that generated these figures in much more depth in Chapters 3–6 and Chapter 9.

New work that Dar is fostering involves building better high school athlete recruiting models. For example, each year the team gives scholarships to three students who are wide receiver recruits. For Dar, picking out the best players goes beyond simple

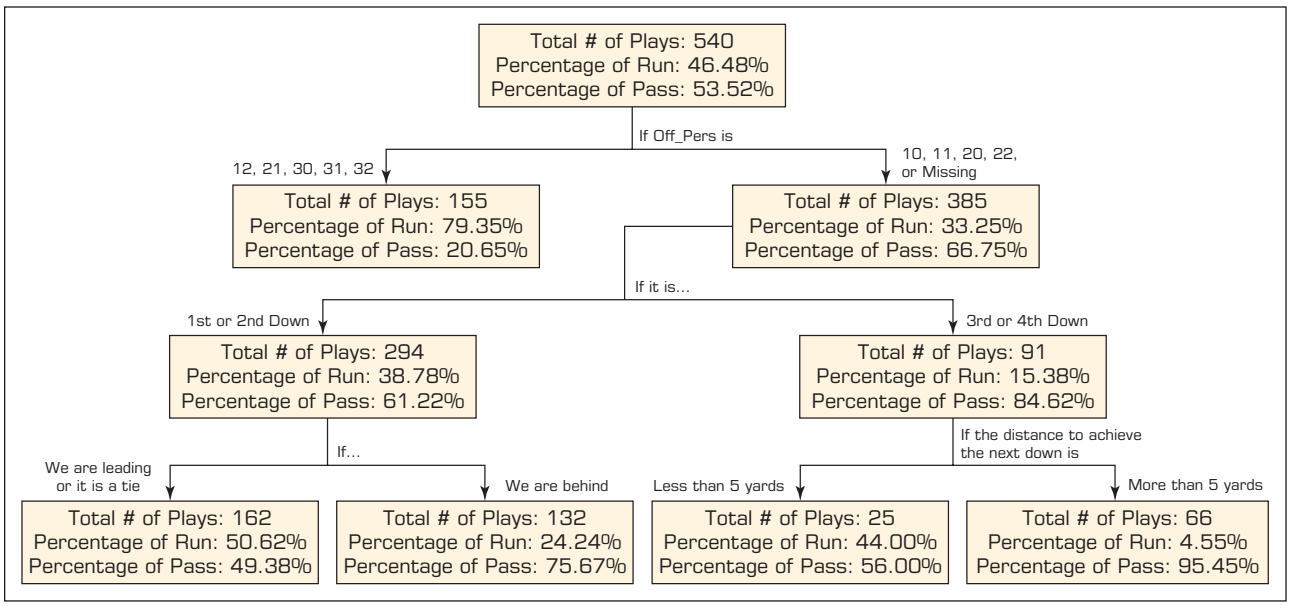


FIGURE 1.12 Cascaded Decision Tree for Run or Pass Plays. *Source:* Contributed by Dr. Dave Schrader, who retired after 24 years in advanced development and marketing at Teradata. He has remained on the Board of Advisors of the Teradata University Network, where he spends his retirement helping students and faculty learn more about sports analytics. Graphics by Peter Liang and Jacob Pearson, graduate students at Oklahoma State University, as part of a student project in the spring of 2016 in Prof. Ramesh Sharda's class under Dr. Dave Schrader's coaching.

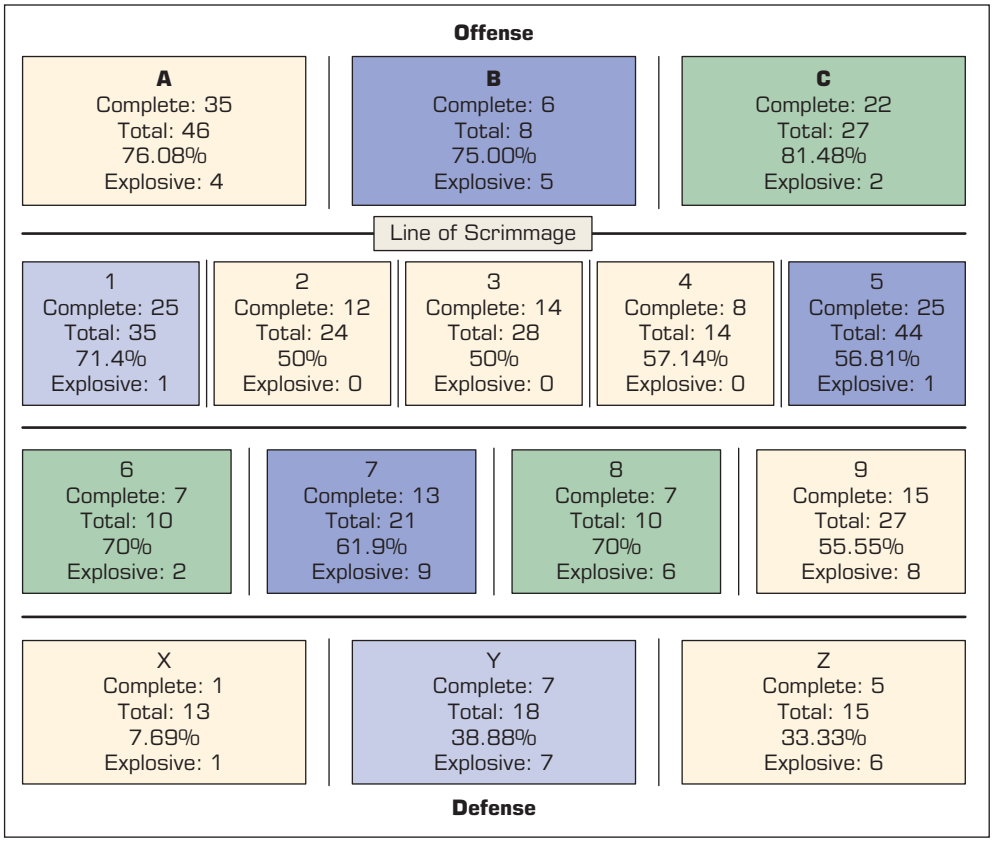


FIGURE 1.13 Heat Map Zone Analysis for Passing Plays. *Source:* Contributed by Dr. Dave Schrader, who retired after 24 years in advanced development and marketing at Teradata. He has remained on the Board of Advisors of the Teradata University Network, where he spends his retirement helping students and faculty learn more about sports analytics. Graphics by Peter Liang and Jacob Pearson, graduate students at Oklahoma State University, as part of a student project in the spring of 2016 in Prof. Ramesh Sharda's class under Dr. Dave Schrader's coaching.

measures like how fast athletes run, how high they jump, or how long their arms are to newer criteria like how quickly they can rotate their heads to catch a pass, what kinds of reaction times they exhibit to multiple stimuli, and how accurately they run pass routes. Some of her ideas illustrating these concepts appear on the TUN Web site; look for the Business Scenario Investigation (2015) "The Case of Precision Football."

WHAT CAN WE LEARN FROM THESE EXAMPLES? Beyond the front-office business analysts, the coaches, trainers, and performance experts, there are many other people in sports who use data, ranging from golf groundskeepers who measure soil and turf conditions for PGA tournaments to baseball and basketball referees who are rated on the correct and incorrect calls they make. In fact, it is hard to find an area of sports that is *not* being impacted by the availability of more data, especially from sensors.

Skills you will learn in this book for business analytics will apply to sports. If you want to dig deeper into this area, we encourage you to look at the Sports Analytics section of the TUN, a free resource for students and faculty. On its Web site, you will find descriptions of what to read to find out more about sports analytics, compilations of places where you can find publically available data sets for analysis, as well as examples

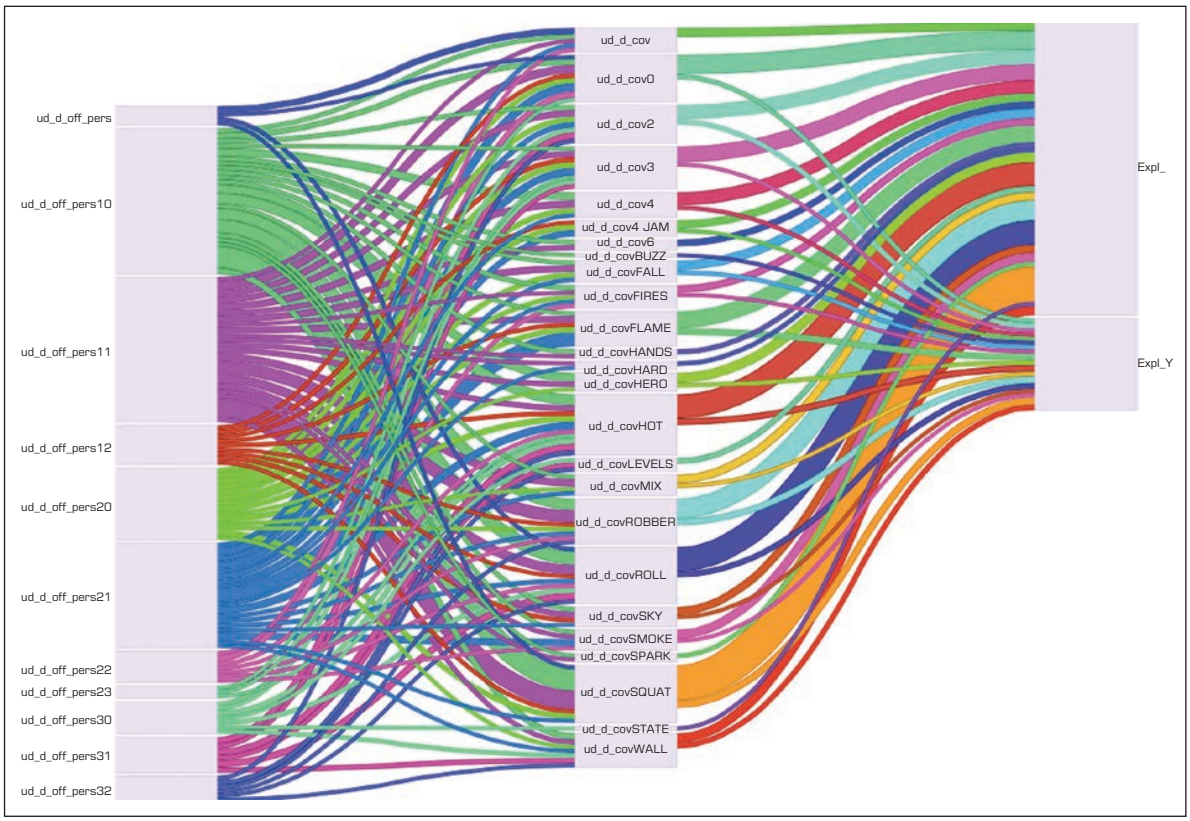


FIGURE 1.14 **Time-Series Analysis of Explosive Plays.**

of student projects in sports analytics and interviews of sports professionals who use data and analytics to do their jobs. Good luck learning analytics!

Analytics Applications in Healthcare—Humana Examples

Although healthcare analytics span a wide variety of applications from prevention to diagnosis to efficient operations and fraud prevention, we focus on some applications that have been developed at a major health insurance company in the United States, Humana. According to its Web site, "The company's strategy integrates care delivery, the member experience, and clinical and consumer insights to encourage engagement, behavior change, proactive clinical outreach and wellness. . . ." Achieving these strategic goals includes significant investments in information technology in general and analytics in particular. Brian LeClaire is senior vice president and CIO of Humana. He has a PhD in MIS from Oklahoma State University. He has championed analytics as a competitive differentiator at Humana—including cosponsoring the creation of a center for excellence in analytics. He described the following projects as examples of Humana's analytics initiatives, led by Humana's chief clinical analytics officer, Vipin Gopal.

Humana Example 1: Preventing Falls in a Senior Population—An Analytic Approach

Accidental falls are a major health risk for adults age 65 years and older with one-third experiencing a fall every year.[1] The costs of falls pose a significant strain on the U.S. healthcare system; the direct costs of falls were estimated at $34 billion in 2013 alone.[1]

With the percent of seniors in the U.S. population on the rise, falls and associated costs are anticipated to increase. According to the Centers for Disease Control and Prevention (CDC), "Falls are a public health problem that is largely preventable" (**www.cdc.gov/homeandrecreationalsafety/falls/adultfalls.html**).[1] Falls are also the leading factor for both fatal and nonfatal injuries in older adults with injurious falls increasing the risk of disability by up to 50 percent (Gill et al., 2013).[2] Humana is the nation's second-largest provider of Medicare Advantage benefits with approximately 3.2 million members, most of whom are seniors. Keeping its senior members well and helping them live safely at their homes is a key business objective of which prevention of falls is an important component. However, no rigorous methodology was available to identify individuals most likely to fall, for whom falls prevention efforts would be beneficial. Unlike chronic medical conditions such as diabetes and cancer, a fall is not a well-defined medical condition. In addition, falls are usually underreported in claims data as physicians typically tend to code the consequence of a fall such as fractures and dislocations. Although many clinically administered assessments to identify fallers exist, they have limited reach and lack sufficient predictive power (Gates et al., 2008).[3] As such, there is a need for a prospective and accurate method to identify individuals at greatest risk of falling so that they can be proactively managed for fall prevention. The Humana analytics team undertook the development of a Falls Predictive Model in this context. It is the first comprehensive PM reported that utilizes administrative medical and pharmacy claims, clinical data, temporal clinical patterns, consumer information, and other data to identify individuals at high risk of falling over a time horizon.

Today, the Falls PM is central to Humana's ability to identify seniors who could benefit from fall mitigation interventions. An initial proof-of-concept with Humana consumers, representing the top 2 percent of those at the highest risk of falling, demonstrated that the consumers had increased utilization of physical therapy services, indicating consumers are taking active steps to reduce their risk for falls. A second initiative utilizes the Falls PM to identify high-risk individuals for remote monitoring programs. Using the PM, Humana was able to identify 20,000 consumers at a high risk of falling who benefited from this program. Identified consumers wear a device that detects falls and alerts a 24/7 service for immediate assistance.

This work was recognized by the Analytics Leadership Award by Indiana University Kelly School of Business in 2015, for innovative adoption of analytics in a business environment.

Contributors: Harpreet Singh, PhD; Vipin Gopal, PhD; Philip Painter, MD.

Humana Example 2: Humana's Bold Goal—Application of Analytics to Define the Right Metrics

In 2014, Humana, Inc. announced its organization's Bold Goal to improve the health of the communities it serves by 20 percent by 2020 by making it easy for people to achieve their best health. The communities that Humana serves can be defined in many ways, including geographically (state, city, neighborhood), by product (Medicare Advantage, employer-based plans, individually purchased), or by clinical profile (priority conditions including diabetes, hypertension, congestive heart failure [CHF], coronary artery disease [CAD], chronic obstructive pulmonary disease [COPD], or depression). Understanding the health of these communities and how they track over time is critical not only for the evaluation of the goal, but also in crafting strategies to improve the health of the whole membership in its entirety.

A challenge before the analytics organization was to identify a metric that captures the essence of the Bold Goal. Objectively measured traditional health insurance

metrics such as hospital admissions or emergency room visits per 1,000 persons would not capture the spirit of this new mission. The goal was to identify a metric that captures health and its improvement in a community and was relevant to Humana as a business. Through rigorous analytic evaluations, Humana eventually selected "Healthy Days," a four-question, quality-of-life questionnaire originally developed by the CDC to track and measure Humana's overall progress toward the Bold Goal.

It was critical to make sure that the selected metric was highly correlated to health and business metrics so that any improvement in Healthy Days resulted in improved health and better business results. Some examples of how "Healthy Days" is correlated to metrics of interest include the following:

- Individuals with more unhealthy days (UHDs) exhibit higher utilization and cost patterns. For a five-day increase in UHDs, there are (1) an $82 increase in average monthly medical and pharmacy costs, (2) an increase of 52 inpatient admits per 1,000 patients, and (3) a 0.28-day increase in average length of stay (Havens, Peña, Slabaugh, Cordier, Renda, & Gopal, 2015).[1]
- Individuals who exhibit healthy behaviors and have their chronic conditions well managed have fewer UHDs. For example, when we look at individuals with diabetes, UHDs are lower if they obtained an LDL screening (−4.3 UHDs) or a diabetic eye exam (−2.3 UHDs). Likewise, if they have controlled blood sugar levels measured by HbA1C (−1.8 UHDs) or LDL levels (−1.3 UHDs) (Havens, Slabaugh, Peña, Haugh, & Gopal 2015).[2]
- Individuals with chronic conditions have more UHDs than those who do not have (1) CHF (16.9 UHDs), (2) CAD (14.4 UHDs), (3) hypertension (13.3 UHDs), (4) diabetes (14.7 UHDs), (5) COPD (17.4 UHDs), or (6) depression (22.4 UHDs) (Havens, Peña, Slabaugh et al., 2015; Chiguluri, Guthikonda, Slabaugh, Havens, Peña, & Cordier, 2015; Cordier et al., 2015).[1,3,4]

Humana has since adopted Healthy Days as their metric for the measurement of progress toward Bold Goal (Humana, **http://populationhealth.humana.com/wp-content/uploads/2016/05/BoldGoal2016ProgressReport_1.pdf**).[5]

Contributors: Tristan Cordier, MPH; Gil Haugh, MS; Jonathan Peña, MS; Eriv Havens, MS; Vipin Gopal, PhD.

Humana Example 3: Predictive Models to Identify the Highest Risk Membership in a Health Insurer

The 80/20 rule generally applies in healthcare; that is, roughly 20 percent of consumers account for 80 percent of healthcare resources due to their deteriorating health and chronic conditions. Health insurers like Humana have typically enrolled the highest-risk enrollees in clinical and disease management programs to help manage the chronic conditions the members have.

Identification of the correct members is critical for this exercise, and in the recent years, PMs have been developed to identify enrollees with high future risk. Many of these PMs were developed with heavy reliance on medical claims data, which results from the medical services that the enrollees use. Because of the lag that exists in submitting and processing claims data, there is a corresponding lag in identification of high-risk members for clinical program enrollment. This issue is especially relevant when new members join a health insurer as they would not have a claims history with an insurer. A claims-based PM could take on average of 9–12 months after enrollment of new members to identify them for referral to clinical programs.

In the early part of this decade, Humana attracted large numbers of new members in its Medicare Advantage products and needed a better way to clinically manage this

membership. As such, it became extremely important that a different analytic approach be developed to rapidly and accurately identify high-risk new members for clinical management, to keep this group healthy and costs down.

Humana's Clinical Analytics team developed the New Member Predictive Model (NMPM) that would quickly identify at-risk individuals soon after their new plan enrollments with Humana rather than waiting for sufficient claim history to become available for compiling clinical profiles and predicting future health risk. Designed to address the unique challenges associated with new members, NMPM developed a novel approach that leveraged and integrated broader data sets beyond medical claims data such as self-reported health risk assessment data and early indicators from pharmacy data, employed advanced data mining techniques for pattern discovery, and scored every Medicare Advantage (MA, a specific insurance plan) consumer daily based on the most recent data Humana has to date. The model was deployed with a cross-functional team of analytics, IT, and operations to ensure seamless operational and business integration.

Since NMPM was implemented in January 2013, it has been rapidly identifying high-risk new members for enrollment in Humana's clinical programs. The positive outcomes achieved through this model have been highlighted in multiple senior leader communications from Humana. In the first quarter 2013 earnings release presentation to investors, Bruce Broussard, CEO of Humana, stated the significance of "improvement in new member PMs and clinical assessment processes," which resulted in 31,000 new members enrolled in clinical programs, compared to 4,000 in the same period a year earlier, a 675 percent increase. In addition to the increased volume of clinical program enrollments, outcome studies showed that the newly enrolled consumers identified by NMPM were also referred to clinical programs sooner with over 50 percent of the referrals identified within the first three months after new MA plan enrollments. The consumers identified also participated at a higher rate and had longer tenure in the programs.

Contributors: Sandy Chiu, MS; Vipin Gopal, PhD.

These examples illustrate how an organization explores and implements analytics applications to meet its strategic goals. You will see several other examples of healthcare applications throughout various chapters in the book.

ANALYTICS IN THE RETAIL VALUE CHAIN The retail sector is where you would perhaps see the most applications of analytics. This is the domain where the volumes are large but the margins are usually thin. Customers' tastes and preferences change frequently. Physical and online stores face many challenges to succeed. And market dominance at one time does not guarantee continued success. So investing in learning about your suppliers, customers, employees, and all the stakeholders that enable a retail value chain to succeed and using that information to make better decisions has been a goal of the analytics industry for a long time. Even casual readers of analytics probably know about Amazon's enormous investments in analytics to power their value chain. Similarly, Walmart, Target, and other major retailers have invested millions of dollars in analytics for their supply chains. Most of the analytics technology and service providers have a major presence in retail analytics. Coverage of even a small portion of those applications to achieve our exposure goal could fill a whole book. So this section highlights just a few potential applications. Most of these have been fielded by many retailers and are available through many technology providers, so in this section, we will take a more general view rather than point to specific cases. This general view has been proposed by Abhishek Rathi, CEO of **vCreaTek.com**. vCreaTek, LLC is a boutique analytics software and service company that has offices in India, the United States, the United Arab Emirates (UAE), and Belgium. The company develops applications in multiple domains, but retail analytics is one of its key focus areas.

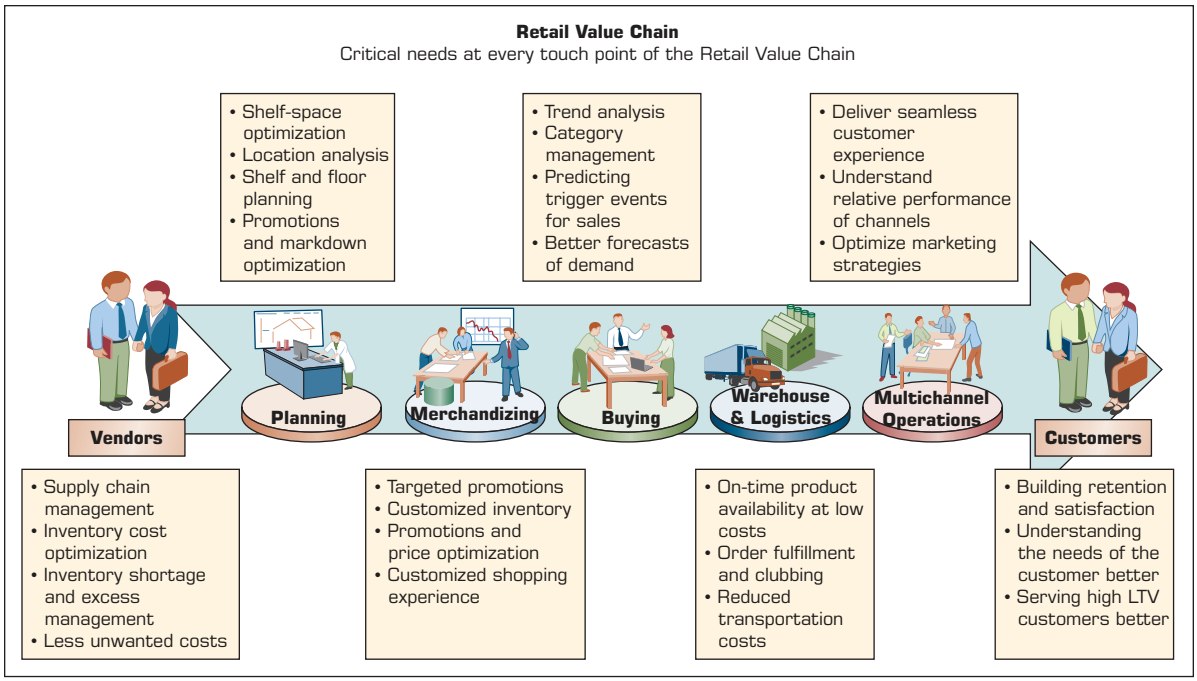


FIGURE 1.15 Example of Analytics Applications in a Retail Value Chain. *Source:* Contributed by Abhishek Rathi, CEO, **vCreaTek.com**.

Figure 1.15 highlights selected components of a retail value chain. It starts with suppliers and concludes with customers but illustrates many intermediate strategic and operational planning decision points where analytics—descriptive, predictive, or prescriptive—can play a role in making better data-driven decisions. Table 1.1 also illustrates some of the important areas of analytics applications, examples of key questions that can be answered through analytics, and of course, the potential business value derived from fielding such analytics. Some examples are discussed next.

An online retail site usually knows its customer as soon as the customer signs in, and thus they can offer customized pages/offerings to enhance the experience. For any retail store, knowing its customer at the store entrance is still a huge challenge. By combining the video analytics and information/badge issued through its loyalty program, the store may be able to identify the customer at the entrance itself and thus enable an extra opportunity for a cross-selling or up-selling. Moreover, a personalized shopping experience can be provided with more customized engagement during the customer's time in the store.

Store retailers invest lots of money in attractive window displays, promotional events, customized graphics, store decorations, printed ads, and banners. To discern the effectiveness of these marketing methods, the team can use shopper analytics by observing closed-circuit television (CCTV) images to figure out the demographic details of the in-store foot traffic. The CCTV images can be analyzed using advanced algorithms to derive demographic details such as age, gender, and mood of the person browsing through the store.

Further, the customer's in-store movement data when combined with shelf layout and planogram can give more insight to the store manager to identify the hot-selling/profitable areas within the store. Moreover, the store manager also can use this information to plan the workforce allocation for those areas for peak periods.

TABLE 1.1 Examples of Analytics Applications in the Retail Value Chain

Analytic Application	Business Question	Business Value
Inventory Optimization	1. Which products have high demand? 2. Which products are slow moving or becoming obsolete?	1. Forecast the consumption of fast-moving products and order them with sufficient inventory to avoid a stock out scenario. 2. Perform fast inventory turnover of slow-moving products by combining them with one in high demand.
Price Elasticity	1. How much net margin do I have on the product? 2. How much discount can I give on this product?	1. Markdown prices for each product can be optimized to reduce the margin dollar loss. 2. Optimized price for the bundle of products is identified to save the margin dollar.
Market-Basket Analysis	1. What products should I combine to create a bundle offer? 2. Should I combine products based on slow-moving and fast-moving characteristics? 3. Should I create a bundle from the same category or a different category line?	1. The affinity analysis identifies the hidden correlations between the products, which can help in following values: a. Strategize the product bundle offering based on focus on inventory or margin. b. Increase cross-selling or up-selling by creating bundle from different categories or the same categories, respectively.
Shopper Insight	1. Which customer is buying what product at what location?	1. By customer segmentation, the business owner can create personalized offers resulting in better customer experience and retention of the customer.
Customer Churn Analysis	1. Who are the customers who will not return? 2. How much business will I lose? 3. How can I retain the customers? 4. What demography of customer is my loyal customer?	1. Businesses can identify the customer and product relationships that are not working and show high churn. Thus, they can have better focus on product quality and the reason for that churn. 2. Based on the customer lifetime value (LTV), the business can do targeted marketing resulting in retention of the customer.
Channel Analysis	1. Which channel has lower customer acquisition cost? 2. Which channel has better customer retention? 3. Which channel is more profitable?	1. Marketing budget can be optimized based on insight for better return on investment.
New Store Analysis	1. What location should I open? 2. What and how much opening inventory should I keep?	1. Best practices of other locations and channels can be used to get a jump-start. 2. Comparison with competitor data can help to create a differentiator to attract the new customers.
Store Layout	1. How should I do store layout for better topline? 2. How can I increase my in-store customer experience?	1. Understand the association of products to decide store layout and better alignment with customer needs. 2. Workforce deployment can be planned for better customer interactivity and thus satisfying customer experience.
Video Analytics	1. What demography is entering the store during the peak period of sales? 2. How can I identify a customer with high LTV at the store entrance so that a better personalized experience can be provided to this customer?	1. In-store promotions and events can be planned based on the demography of incoming traffic. 2. Targeted customer engagement and instant discount enhances the customer experience resulting in higher retention.

Market-basket analysis has commonly been used by the category managers to push the sale of slowly moving stock keeping units (SKUs). By using advanced analytics of data available, the product affinity can be identified at the lowest level of SKU to drive better returns on investments (ROIs) on the bundle offers. Moreover, by using price elasticity techniques, the markdown or optimum price of the bundle offer can also be deduced, thus reducing any loss in the profit margin.

Thus, by using data analytics, a retailer can not only get information on its current operations but can also get further insight to increase the revenue and decrease the operational cost for higher profit. A fairly comprehensive list of current and potential retail analytics applications that a major retailer such as Amazon could use is proposed by a blogger at Data Science Central. That list is available at **www.datasciencecentral.com/profiles/blogs/20-data-science-systems-used-by-amazon-to-operate-its-business**. As noted earlier, there are too many examples of these opportunities to list here, but you will see many examples of such applications throughout the book.

IMAGE ANALYTICS As seen in this section, analytics techniques are being applied to many diverse industries and data. An area of particular growth has been analysis of visual images. Advances in image capturing through high-resolution cameras, storage capabilities, and deep learning algorithms have enabled very interesting analyses. Satellite data have often proven their utility in many different fields. The benefits of satellite data at high resolution and in different forms of imagery including multi-spectral are significant to scientists who need to regularly monitor global change, land usage, and weather. In fact, by combining the satellite imagery and other data including information on social media, government filings, and so on, one can surmise business planning activities, traffic patterns, changes in parking lots or open spaces. Companies, government agencies, and non-governmental organizations (NGOs) have invested in satellites to try to image the whole globe every day so that daily changes can be tracked at any location and the information can be used for forecasting. In the last few months, many interesting examples of more reliable and advanced forecasts have been reported. Indeed, this activity is being led by different industries across the globe, and has added a term to Big Data called *Alternative Data*. Here are a few examples from Tartar et al. (2018). We will see more in Chapter 9 when we study Big Data.

- World Bank researchers used satellite data to propose strategic recommendations for urban planners and officials from developing nations. This analysis arose due to the recent natural disaster where at least 400 people died in Freetown, Sierra Leone. Researchers clearly demonstrated that Freetown and some other developing cities lacked systematic planning of their infrastructure that resulted in the loss of life. The bank researchers are using satellite imagery now to make critical decisions regarding risk-prone urban areas.
- EarthCast provides accurate weather updates for a large commercial U.S. airline based on the data it pulls from a constellation of 60 government-operated satellites combined with ground and aircraft-based sensors, tracking almost anything from lightning to turbulence. It has even developed the capability to map out conditions along a flight path and provides customized forecasts for everything from hot air balloons to drones.
- Imazon started using satellite data to develop a picture of close real-time information on Amazon deforestation. It uses advanced optical and infrared imagery that has led to identifying illegal sawmills. Imazon is now focused more on getting data to local governments through its "green municipalities" program that trains officials to identify and curb deforestation.

- The Indonesian government teamed up with international nonprofit Global Fishing Watch, which processes satellite extracted information on ship movement to spot where and when vessels are fishing illegally (Emmert, 2018). This initiative delivered instant results: Government revenue from fishing went up by 129 percent in 2017 compared to 2014. It is expected that by next decade, the organization would track vessels that are responsible for 75 percent of the world's catch.

These examples illustrate just a sample of ways that satellite data can be combined with analytics to generate new insights. In anticipation of the coming era of abundant earth observations from satellites, scientists and communities must put some thought into recognizing key applications and key scientific issues for the betterment of society. Although such concerns will eventually be resolved by policymakers, what is clear is that new and interesting ways of combining satellite data and many other data sources is spawning a new crop of analytics companies.

Such image analysis is not limited to satellite images. Cameras mounted on drones and traffic lights on every conveyable pole in buildings and streets provide the ability to capture images from just a few feet high. Analysis of these images coupled with facial recognition technologies is enabling all kinds of new applications from customer recognition to governments' ability to track all subjects of interest. See Yue (2017) as an example. Applications of this type are leading to much discussion on privacy issues. In Application Case 1.7, we learn about a more benevolent application of image analytics where the images are captured by a phone and a mobile application provides immediate value to the user of the app.

Application Case 1.7 Image Analysis Helps Estimate Plant Cover

Estimating how much ground is covered by green vegetation is important in analysis of a forest or even a farm. In case of a forest, such analysis helps users understand how the forest is evolving, its impact on surrounding areas, and even climate. For a farm, similar analysis can help understand likely plant growth and help estimate future crop yields. It is obviously impossible to measure all forest cover manually and is challenging for a farm. The common method is to record images of a forest/farm and then analyze these images to estimate the ground cover. Such analysis is expensive to perform visually and is also error prone. Different experts looking at the ground cover might estimate the percentage of ground covering differently. Thus, automated methods to analyze these images and estimate the percentage of ground covered by vegetation are being developed. One such method and an app to make it practical through a mobile phone has been developed at Oklahoma State University by researchers in the Department of Plant and Soil Sciences in partnership with the university's App Center and the Information Technology group within the Division of Agricultural Sciences and Natural Resources.

Canopeo is a free desktop or mobile app that estimates green canopy cover in near real-time from images taken with a smartphone or digital camera. In experiments in corn, wheat, canola, and other crops, Canopeo calculated the percentage of canopy covering dozens to thousands of times faster than existing software without sacrificing accuracy. And unlike other programs, the app can acquire and analyze video images, says Oklahoma State University (OSU) soil physicist, Tyson Ochsner—a feature that should reduce the sampling error associated with canopy cover estimates. "We know that plant cover, plant canopies, can be quite variable in space," says Ochsner, who led the app's development with former doctoral student Andres Patrignani, now a faculty member at Kansas State University. "With Canopeo, you can just turn on your [video] device, start walking across a portion of a field, and get results for every frame of video that you're recording." By using a smartphone or tablet's digital camera, Canopeo users in the field can take photos or videos of green plants, including crops, forages, and turf, and import them to the app, which analyzes each image pixel, classifying them based on its red-green-blue (RGB)

color values. Canopeo analyzes pixels based on a ratio of red to green and blue to green pixels as well as an excess green index. The result is an image where color pixels are converted into black and white with white pixels corresponding to green canopy and black pixels representing the background. Comparison tests showed that Canopeo analyzes images more quickly and just as accurately as two other available software packages.

Developers of Canopeo expect the app to help producers judge when to remove grazing cattle from winter wheat in "dual-purpose" systems where wheat is also harvested for grain. Research by others at OSU found that taking cattle off fields when at least 60 percent green canopy cover remained ensured a good grain yield. "So, Canopeo would be useful for that decision," Patrignani says. He and Ochsner also think the app could find use in turfgrass management; in assessments of crop damage from weather or herbicide drift; as a surrogate for the Normalized Difference Vegetation Index (NDVI) in fertilizer recommendations; and even in UAV-based photos of forests or aquatic systems.

Analysis of images is a growing application area for deep learning as well as many other AI techniques. Chapter 9 includes several examples of image analysis that have spawned another term—alternative data. Applications of alternative data are emerging in many fields. Chapter 6 also highlights some applications. Imagining innovative applications by being exposed to others' ideas is one of the main goals of this book!

Questions for Discussion

1. What is the purpose of knowing how much ground is covered by green foliage on a farm? In a forest?
2. Why would image analysis of foliage through an app be better than a visual check?
3. Explore research papers to understand the underlying algorithmic logic of image analysis. What did you learn?
4. What other applications of image analysis can you think of?

Source: Compiled from A. Patrignani and T. E. Ochsner. (2015). "Canopeo: A Powerful New Tool for Measuring Fractional Green Canopy Cover." *Agronomy Journal, 107*(6), pp. 2312–2320; R. Lollato, A. Patrignani, T. E. Ochsner, A. Rocatelli, P. Tomlinson, & J. T. Edwards. (2015). Improving Grazing Management Using a Smartphone App. **www.bookstore.ksre.ksu.edu/pubs/MF3304.pdf** (accessed October 2018); **http://canopeoapp.com/** (accessed October 2018); Oklahoma State University press releases.

Analytics/data science initiatives are quickly embracing and even merging with new developments in artificial intelligence. The next section provides an overview of artificial intelligence followed by a brief discussion of convergence of the two.

SECTION 1.6 REVIEW QUESTIONS

1. What are three factors that might be part of a PM for season ticket renewals?
2. What are two techniques that football teams can use to do opponent analysis?
3. What other analytics uses can you envision in sports?
4. Why would a health insurance company invest in analytics beyond fraud detection? Why is it in its best interest to predict the likelihood of falls by patients?
5. What other applications similar to prediction of falls can you envision?
6. How would you convince a new health insurance customer to adopt healthier lifestyles (Humana Example 3)?
7. Identify at least three other opportunities for applying analytics in the retail value chain beyond those covered in this section.
8. Which retail stores that you know of employ some of the analytics applications identified in this section?
9. What is a common thread in the examples discussed in image analytics?
10. Can you think of other applications using satellite data along the lines presented in this section?

1.7 ARTIFICIAL INTELLIGENCE OVERVIEW

On September 1, 2017, the first day of the school year in Russia, Vladimir Putin, the Russian President, lectured to over 1,000,000 school children in what is called in Russia the National Open Lesson Day. The televised speech was titled "Russia Focused on the Future." In this presentation, the viewers saw what Russian scientists are achieving in several fields. But, what everyone remembers from this presentation is one sentence: "The country that takes the lead in the sphere of computer-based artificial intelligence will become the ruler of the world."

Putin is not the only one who knows the value of AI. Governments and corporations are spending billions of dollars in a race to become a leader in AI. For example, in July 2017, China unveiled a plan to create an AI industry worth $150 billion to the Chinese economy by 2030 (Metz, 2018). China's Badu Company today employs over 5,000 AI engineers. The Chinese government facilitates research and applications as a national top priority. The accounting firm PricewaterhouseCoopers calculated that AI will add $15.7 trillion to the global economy by 2030 (about 14 percent; see Liberto, 2017). Thus, there is no wonder that AI is clearly the most talked about technology topic in 2018.

What Is Artificial Intelligence?

There are several definitions of what is AI (Chapter 2). The reason is that AI is based on theories from several scientific fields, and it encompasses a wide collection of technologies and applications. So, it may be beneficial to look at some of the characteristics of AI in order to understand what it is. The major goal of AI is to create intelligent machines that can do tasks currently done by people. Ideally, these tasks include reasoning, thinking, learning, and problem solving. AI studies the human thought processes' ability to understand what intelligence is so AI scientists can duplicate the human processes in machines. eMarketer (2017) provides a comprehensive report, describing AI as

- Technology that can learn to do things better over time.
- Technology that can understand human language.
- Technology that can answer questions.

The Major Benefits of AI

Since AI appears in many shapes, it has many benefits. They are listed in Chapter 2. The major benefits are as follows:

- Significant reduction in the cost of performing work. This reduction continues over time while the cost of doing the same work manually increases with time.
- Work can be performed much faster.
- Work is consistent in general, more consistent than human work.
- Increased productivity and profitability as well as a competitive advantage are the major drivers of AI.

The Landscape of AI

There are many parts in the landscape (or ecosystem) of AI. We decided to organize them into five groups as illustrated in Figure 1.16. Four of the groups constitute the basis for the fifth one, which is the AI applications. The groups are as follows:

MAJOR TECHNOLOGIES Here we elected to include machine learning (Chapter 5), deep learning (Chapter 6), and intelligent agents (Chapter 2).

KNOWLEDGE-BASED TECHNOLOGIES (all covered in Chapter 12) Topics covered are expert systems, recommendation engines, chatbots, virtual personal assistants, and robo-advisors.

BIOMETRIC-RELATED TECHNOLOGIES This includes natural language processing (understanding and generation, machine vision and scene and image recognition and voice and other biometric recognition (Chapter 6).

SUPPORT THEORIES, TOOLS, AND PLATFORMS Academic disciplines include computer science, cognitive science, control theory, linguistics, mathematics, neuroscience, philosophy, psychology, and statistics.

Devices and methods include sensors, augmented reality, context awareness, logic, gestural computing collaborative filtering, content recognition, neural networks, data mining, humanoid theories, case-based reasoning, predictive application programming interfaces (APIs), knowledge management, fuzzy logic, genetic algorithm, bin data, and much more.

TOOLS AND PLATFORMS These are available from IBM, Microsoft, Nvidia, and several hundred vendors specializing in the various aspects of AI.

AI APPLICATIONS There are several hundred or may be thousands of them. We provide here only a sample:

Smart cities, smart homes, autonomous vehicles (Chapter 13), automatic decisions (Chapter 2), language translation, robotics (Chapter 10), fraud detection, security protection, content screening, prediction, personalized services, and more. Applications are in all business areas (Chapter 2), and in almost any other area ranging from medicine and healthcare to transportation and education.

Note: Lists of all these are available at Faggela (2018) and Jacquet (2017). Also see Wikipedia, "Outline of artificial intelligence," and a list of "AI projects" (several hundred items.)

In Application Case 1.8, we describe how several of these technologies are combined in improving security and in expediting the processing of passengers in airports.

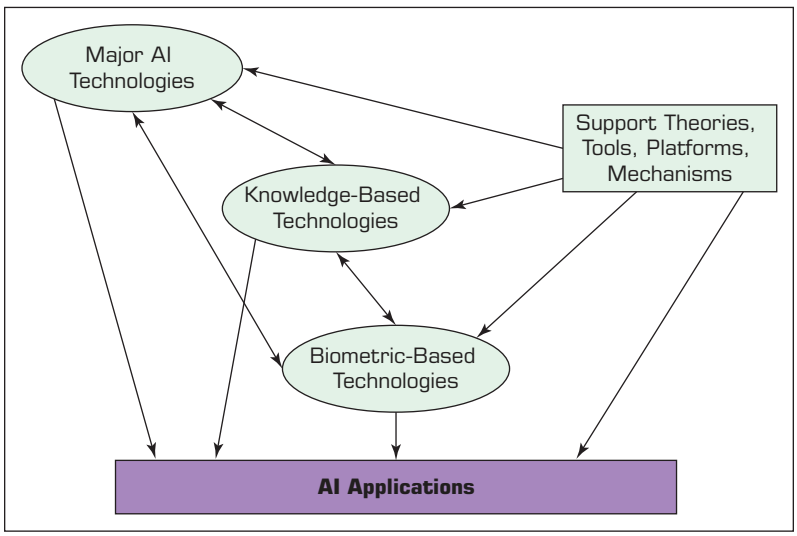


FIGURE 1.16 **The Landscape (Ecosystem) of AI.** *Source:* Drawn by E. Turban.

Application Case 1.8

AI Increases Passengers' Comfort and Security in Airports and Borders

We may not like the security lines at airports or the idea that terrorists may board our plane or enter our country. Several AI devices are designed to minimize these possibilities.

1. *Facial recognition at airports.* Jet Blue is experimenting with facial-recognition technology (a kind of machine vision to match travelers' faces against prestored photos, such as passport, driver's license). This will eliminate the need for boarding passes and increase security. The match is of high quality. The technology pioneered by British Airways is used by Delta, KLM, and other airlines using similar technologies for self-checking of bags. Similar technology is used by the U.S. Immigration and Customs Enforcement agency where people's photos taken at arrivals are matched against the database of photos and other documents.
2. *China's system.* The major airports in China are using a system similar to that of Jet Blue, using *facial recognition* for verifying the identity of passengers. The idea is to eliminate boarding passes and expedite the flow of boarding. The system is also used to recognize airport employees entering restricted areas.
3. *Using bots.* Several airports (e.g., New York, Beijing) offer conversational bots (Chapter 12) to provide travelers with airport guidance. Bots provide also information about customs and immigration services.
4. *Spotting liars at airport.* This application is emerging to help immigration services to vet passengers at airports and land entry borders. With increased security, both immigration and airline personal may need to query passengers. Here is the solution that can be economically used to query all passengers at high speed, so there will be short waiting lines. This emerging system is called Automated Virtual Agent for Truth Assessments in Real Time (AVATAR). The essentials of the system are as follows:
   - **a.** AVATAR is a bot in which you first scan your passport.
   - **b.** AVATAR asks you a few questions. Several AI technologies are used in this project, such as AI, Big Data analytics, the "Cloud," robotics, machine learning, machine vision, and bots.
   - **c.** You answer the questions.
   - **d.** AVATAR's sensors and other AI technologies collect data from your body, such as voice variability, facial expression (e.g., muscle engagement), eyes' position and movements, mouth movements, and body posture. Researchers feel that it takes less effort to tell the truth than to die, so researchers compared the answers to routine questions.

The machine then will flag suspects for further investigation. The machine is already in use by immigration agents in several countries.

Sources: Condensed from Thibodeaux, W. (2017, June 29). "This Artificial Intelligence Kiosk Is Designed to Spot Liars at Airports." **Inc.com**.; Silk, R. (2017, November). "Biometrics: Facial Recognition Tech Coming to an Airport Near You." *Travel Weekly, 21.*

Questions for Case 1.8

1. List the benefits of AI devices to travelers.
2. List the benefits to governments and airline companies.
3. Relate this case to machine vision and other AI tools that deal with people's biometrics.

NARROW (WEAK) VERSUS GENERAL (STRONG) AI The AI field can be divided into two major categories of applications: narrow (or weak) and general (or strong).

A Narrow AI Field Focuses on One Narrow Field (Domain). Well-known examples of this are SIRI and Alexa (Chapter 12) that, at least in their early years of life, operated in limited, predefined areas. As time has passed, they have become more general, acquiring additional knowledge. Most expert systems (Chapter 12) were operating in fairly narrow domains. If you notice, when you converse with an automated call center, the computer

(which is usually based on some AI technology) is not too intelligent. But, it is getting "smarter" with time. Speech recognition allows computers to convert sound to text with great accuracy. Similarly, computer vision is improving, recognizes objects, classifies them, and even understands their movements. In sum, there are millions of narrow AI applications, and the technology is improving every day. However, AI is not strong enough yet because it does not exhibit the true capabilities of human intelligence (Chapter 2).

GENERAL (STRONG) AI To exhibit real intelligence, machines need to perform the full range of human cognitive capabilities. Computers can have some cognitive capabilities (e.g., some reasoning and problem solving) as will be shown in Chapter 6 on cognitive computing.

The difference between the two classes of AI is getting smaller as AI is getting smarter. Ideally, strong AI will be able to replicate humans. But true intelligence is happening only in narrow domains, such as game playing, medical diagnosis, and equipment failure diagnosis.

Some feel that we never will be able to build a truly strong AI machine. Others think differently; see the debate in Section 14.9. The following is an example of a strong AI bot in a narrow domain.

Example 3: AI Makes Coke Vending Machine Smarter

If you live in Australia or New Zealand and you are near a Coca-Cola vending machine, you can order a can or a bottle of the soft drink using your smartphone. The machines are cloud connected, which means you can order the Coke from any place in the world, not only for yourself but also for any friend who is near a vending machine in Australia or New Zealand. See Olshansky (2017).

In addition, the company can adjust pricing remotely, offer promotions, and collect inventory data so that restocking can be made. Converting existing machines to AI-enabled takes about 1 hour each.

Wait a minute, what if something goes wrong? No problem, you can chat with Coca-Cola's bot via Facebook Messenger (Chapter 12).

The Three Flavors of AI Decisions

Staff (2017) divided the capabilities of AI systems into three levels: assisted, autonomous, and augmented.

ASSISTED INTELLIGENCE This is equivalent mostly to the weak AI, which works only in narrow domains. It requires clearly defined inputs and outputs. Examples are some monitoring systems and low-level virtual personal assistants (Chapter 12). Our cars are full of such monitoring systems that give us alerts. Similarly, there are many healthcare applications (monitoring, diagnosis).

Autonomous AI

These systems are in the realm of the strong AI but in very narrow domain. Eventually, the computer will take over. Machines will act as experts and have absolute decision-making power. Pure robo-advisors (Chapter 12) are examples of such machines. Autonomous vehicles and robots that can fix themselves are also good examples.

AUGMENTED INTELLIGENCE Most of the existing AI applications are between assisted and autonomous and/are referred to as **augmented intelligence** (or intelligence augmentation). The technology focuses on augmenting computer abilities to extend human cognitive abilities (see Chapter 6 on cognitive computing), resulting in high performance as described in Technology Insights 1.1.

TECHNOLOGY INSIGHTS 1.1 Augmented Intelligence

The idea of combining the performance of people and machines is not new. Here we combine (augmenting) human capabilities with powerful machine intelligence. That is, not replacing people which is done by autonomous AI, but extending human cognitive abilities. The result is the ability to solve complex human problems as in the opening vignette to this chapter. The computers enabled people to solve problems that were unsolved before. Padmanabhan (2018) distinguishes the following differences between traditional and augmented AI:

1. Augmented machines extend rather than replace human decision making, and they facilitate creativity.
2. Augmentation excels in solving complex human and industry problems in specific domains in contrast with strong, general AI.
3. In contrast with a "black box" model of some AI and analytics, augmented intelligence provides insights and recommendations, including explanations.
4. In addition, augmentation technology can offer new solutions by combining existing and discovered information in contrast with assisted AI, which identifies problems or symptoms and suggests predetermined solutions.

Padmanabhan (2018) and many others believe that at the moment, augmented AI is the best option to move toward the transformation of the AI world.

In contrast with autonomous AI, which describes machines with a wide range of cognitive abilities (e.g., driverless vehicles), augmented intelligence has only a few cognitive abilities.

Examples of Augmented Intelligence

Staff (2017) provides the following examples:

- **Cybercrime fighting.** AI can identify forthcoming attacks and suggest solutions.
- **e-Commerce decisions.** Marketing tools make testing 100 times faster and adapt the layout and response functions of a Web site to users. The machines make recommendations and the marketers can accept or reject them.
- **High-frequency stock market trading.** This is done either completely autonomously or in some cases with control and calibration by humans.

DISCUSSION QUESTIONS

1. What is the basic premise of augmented intelligence?
2. List the major differences between augmented intelligence and traditional AI.
3. What are some benefits of augmented intelligence?
4. How does technology relate to cognitive computing?

Societal Impacts

Much talk is on the topics of AI and productivity, speed, and cost reduction. In a national conference hosted by Gartner, the famous IT consulting company, nearly half of 3,000 participating U.S. CIOs reported plans to deploy AI-now (Weldon, 2018). Industry cannot ignore the potential benefits of AI, especially its increased productivity gains, cost reduction and quality, and speed. Conference participants there talked about strategy and implementation (Chapter 14). It seems that every company is at least involved in piloting and experimentation AI. However, in all this excitement, we should not neglect the societal impacts. Many of these are positive, some are negative, and most are unknown. A comprehensive discussion is provided in Chapter 14. Here we provide three examples of AI impacts.

IMPACT ON AGRICULTURE A major impact of AI will be on agriculture. One major anticipated result is to provide more food, especially in third world countries. Here are a few examples:

- According to Lopez (2017), using AI and robots can help farmers produce 70 percent more food by 2050. This increase is a result of higher productivity of farm equipment boosted by IoT (see opening vignette to Chapter 13) and a reduced cost of producing food. (Today only 10 percent of a family's budget is spent on food versus. 17.5 percent in 1960).
- Machine vision helps in improved planting and harvesting. Also, AI helps to pick good kernels of grain.
- AI will help to improve the nutrition of food.
- AI will reduce the cost of food processing.
- Driverless tractors are already being experimented with.
- Robots know how to pick fruits and to plant vegetables can solve the shortage of farm workers.
- Crop yields are continuously increasing in India and other countries.
- Pest control improves. For example, AI can predict pest attacks, facilitating planning.
- Weather conditions are monitored by satellites. AI algorithms tell farmers when to plant and/or harvest.

The list can go on and on. For countries such as India and Bangladesh, these activities will critically improve the life of farmers. All in all, AI will help farmers make better decisions. For a Bangladesh case, see PE Report (2017). See **alsonews.microsoft.com/en-in/features/ai-agriculture-icrisat-upl-india/**.

Note: AI can help hungry pets too. A food and water dispenser, called Catspad, is available in the United Kingdom for about US $470 You need to put an ID tag on your pet (only cats and small dogs). The dispenser knows which pet comes to the food and dispenses the type and amount of appropriate food. In addition, sensors (Chapter 13) can tell you how much food each pet ate. You will also be notified if water needs to be added. Interested? See Deahl (2018) for details.

INTELLIGENT SYSTEMS CONTRIBUTION TO HEALTH AND MEDICAL CARE Intelligent systems provide a major contribution to our health and medical care. New innovations arrive almost any day in some place in the world (governments, research institutions, and corporation-sponsored active medical AI research). Here are some interesting innovations.

- AI excels in disease prediction (e.g., predicting the occurrence of infective diseases one week in advance).
- AI can detect brain bleeds.
- AI can track medication intake, send medical alerts, order medicine refills, and improve prescription compliance.
- Mobile telepresence robots remotely connect physicians and patients.
- NVIDIA's medical imaging supercomputer helps diagnosticians and facilitates cures of diseases.
- Robotics and AI can redesign pharmaceutical supply chains.
- AI predicts cardiovascular risks from retinal images.
- Cancer predictions are made with deep learning, and machine learning performs melanoma diagnosis.
- A virtual personal assistant can assess a patient's mood and feeling by cues provided (e.g., speech gesture or inflection).
- Many portals provide medical information to patients and even surgeons. Adoptive spine IT is an example.

- Aging-based AI center for research on people who are elderly operates in the United States. Similar activities exist in Japan.
- The use of bionic hands and legs is improving with AI.
- Healthcare IT News (2017) describes how AI is solving healthcare problems by using virtual assistants (Chapter 12).

The list can go on and on. Norman (2018) describes the scenario of replacing doctors with intelligent systems.

Note: AI in medicine is recognized as a scientific field with national and international annual conferences. For a comprehensive book on the subject, see Agah (2017).

OTHER SOCIETAL APPLICATIONS There are many AI applications in transportation, utilities, education, social services, and other fields. Some are covered under the topic of smart cities (Chapter 13). AI is used by social media and others to control content including fake news. Finally, how about using technology to eradicate child slavery in the Middle East? See Application Case 1.9.

Application Case 1.9 Robots Took the Job of Camel-Racing Jockeys for Societal Benefits

In several Middle Eastern countries, notably Jordan, Abu Dhabi, and other Gulf nations, racing camels has been a popular activity for generations. The owners of the winning camels can make a huge bonus (up to $1,000,000 for first place). Also, the events are considered cultural and social.

The Problem

For a long time, the racing camels were guided by human jockeys. The lighter the weight of the rider, the better is the chance to win. So the owners of the camels trained children (as young as seven) to be jockeys. Young male children were bought (or kidnapped) from poor families in Sudan, India, Bangladesh, and other poor countries and were trained as child jockeys. In fact, this practice was using child slave labor to race the camels. This practice was used for generations until it was banned in all Middle Eastern countries during 2005–2010. A major factor that resulted in the banning was the utilization of robots.

The Robots' Solution

Racing camels was a tradition for many generations and become a lucrative sport. So, no one wanted to discontinue it. According to Opfer (2016), there was a humanistic reason for using robots to race camels—to save the children. Today, all camel race tracks in the Middle East employ only robots. The robots are tied to the hump of the camels, looking like small jockeys and are remote controlled from cars that drive parallel to the racing camels. The owners can command the camels by voice, and they can also operate a mechanical whip to beat the animals so they will run faster, much like human jockeys do. Note that camels would not run unless they hear the voice of a human or see something that looks like a human on their humps.

The Technology

There is a video camera that shows the people that are in cars driving alongside of the camels, what is going on in real time. The owner can provide voice commands to the camel from the car. A mechanical whip attached to the hump of the camel can be remotely operated to induce the animal.

The Results

The results are astonishing. Not only was the child slavery practice eliminated, but also the speed obtained by the camels increased. After all, the robots used weigh only 6 pounds and do not get tired. To see how this works watch the video at **youtube.com/watch?v=GVeVhWXB7sk** (2:47 min.). To view a complete race, see **youtube.com/watch?v=xFCRhk4GYds** (9:08 min.). You may have

a chance to see the royal family when you go to the track. Finally, you can see more details in **youtube.com/watch?v=C1uYAXJIbYg** (8:08 min.).

Sources: Compiled from C. Chung. (2016, April 4). "Dubai Camel Race Ride-Along." **YouTube.com**. **youtube.com/watch?v=xFCRhk4GYds** (accessed September 2018); P. Boddington. (2017, January 3). "Case Study: Robot Camel Jockeys. Yes, really." *Ethics for Artificial Intelligence*; and L. Slade. (2017, December 21). "Meet the Jordanian Camel Races Using Robot Jockeys." **Sbs.com.au**.

Discussion Questions

1. It is said that the robots eradicated the child slavery. Explain.
2. Why do the owners need to drive by their camels while they are racing?
3. Why not duplicate the technology for horse racing?
4. Summarize ethical aspects of this case (Read Boddington, 2017). Do this exercise after you have read about ethics in Chapter 14.

SECTION 1.7 REVIEW QUESTIONS

1. What are the major characteristics of AI?
2. List the major benefits of AI.
3. What are the major groups in the ecosystem of AI? List the major contents of each.
4. Why is machine learning so important?
5. Differentiate between narrow and general AI.
6. Some say that no AI application is strong. Why?
7. Define *assisted intelligence, augmented intelligence*, and *autonomous intelligence.*
8. What is the difference between traditional AI and augmented intelligence?
9. Relate types of AI to cognitive computing.
10. List five major AI applications for increasing the food supply.
11. List five contributions of AI in medical care.

1.8 CONVERGENCE OF ANALYTICS AND AI

Until now we have presented analytics and AI as two independent entities. But, as illustrated in the opening vignette, these technologies can be combined in solving complex problems. In this section, we discuss the convergence of these techniques and how they complement each other. We also describe the possible addition of other technologies, especially IoT, that enable the solutions to very complex problems.

Major Differences between Analytics and AI

As you recall from Section 1.4, analytics process historical data using *statistical, management science* and other computational tools to describe situations (descriptive analytics), to predict results including forecasting (predictive analytics), and to propose recommendations for solutions to problems (predictive analytics). The emphasis is on the statistical, management science, and other computational tools that help analyze historical data.

AI, on the other hand, also uses different tools, but its major objective is to mimic the manner in which people think, learn, reason, make decisions, and solve problems. The emphasis here is on *knowledge* and *intelligence* as major tools for solving problems rather than relying on computation, which we do in analysis. Furthermore, AI also is dealing with cognitive computing. In reality, this difference is not so clear because in advanced analytic applications, there are situations of using machine learning (an AI

technology), supporting analytics in both prediction and prescription. In this section, we describe the convergence of intelligent technologies.

Why Combine Intelligent Systems?

Both analytics and AI and their different technologies are making useful contributions to many organizations when each is applied by itself. But each does have limitations According to a Gartner study, the chance that business analytics initiatives will not meet the enterprise objectives is 70–80 percent. Namely, at least 70 percent of corporate needs are not fulfilled. In other words, there is only a small chance that business intelligence initiatives will result in organizational excellence. There are several reasons for this situation including:

- Predictive models have unintended effects (see Chapter 14).
- Models must be used ethically, responsibly, and mindfully (Chapter 14). They may not be used this way.
- The results of analytics may be very good for some applications but not for others.
- Models are as good as their input data and assumptions (garbage-in, garbage-out).
- Data could be incomplete. Changing environments can make data obsolete very quickly. Models may be unable to adapt.
- Data that come from people may not be accurate.
- Data collected from different sources can vary in format and quality.

Additional reasons for combining intelligent systems are generic to IT projects, and they are discussed in Section 14.2.

The failure rate of AI initiatives is also high. Some of the reasons are similar to the rate of analytics. However, a major reason is that some AI technologies need a large amount of data, sometimes Big Data. For example, many millions of data items are fed to Alexa every day to increase its knowledge. Without continuous flow of data, there would not be good learning in AI.

The question is whether AI and analytics (and other intelligent systems) can be combined in such a way that there will be synergy for better results.

How Convergence Can Help?

According to Nadav (2017), business intelligence and its analytics answer most of the *why* and *what* questions regarding the sufficiency of problem solving. Adding prescriptive analytics will add more cost but not necessarily better performance. Therefore, the next generation of business intelligence platforms will use AI to automatically locate, visualize, and narrate important things. This can also be used to create automatic alerts and notifications. In addition, machine learning and deep learning can support analytics by conducting pattern recognition and more accurate predictions. AI will help to compare actual performance with the predicted one (see Section 14.6). Machine learning and other AI technologies also provide for constant improvement strategy. Nadav also suggested adding expert opinions via collective intelligence, as presented in Chapter 11.

In the remaining part of this section, we present detailed aspects of convergence of some intelligent systems.

Big Data Is Empowering AI Technologies

Big Data is characterized by its volume, variety, and velocity that exceed the reach of commonly used hardware environments and/or the capabilities of software tools to process data. However, today there are technologies and methods that enable capturing,

cleaning, and analyzing Big Data. These technologies and methods enable companies to make real-time decisions. The convergence with AI and machine learning is a major force in this direction. The availability of new Big Data analytics enables new capabilities in AI technologies that were not possible until recently. According to Bean (2017), Big Data can empower AI due to:

- The new capabilities of processing Big Data at a much reduced cost.
- The availability of large data sets online.
- The scale up of algorithms, including deep learning, is enabling powerful AI capabilities.

MetLife Example: Convergence of AI and Big Data

MetLife is a Canadian-based global insurance company that is known for its use of IT to smooth its operation and increase customer satisfaction. To get the most from technology, the company uses AI that has been enabled by Big Data analysis as follows:

- Tracking incidents and their outcomes has been improved by speech recognition.
- Machine learning indicates pending failures. In addition, handwritten reports made by doctors about people injured or were sick and claims paid by the insurance company are analyzed in seconds by the system.
- Expediting the execution of underwriting policies in property and casualty insurance is done by using both AI and analytics.
- The back-office side of claim processing includes many unstructured data that are incorporated in claims. Part of the analysis includes patients' health data. Machine learning is used to recognize anomalies in reports very quickly.

For more about AI and the insurance business, see Chapter 2. For more on the convergence of Big Data and AI in general and at MetLife, see Bean (2017).

The Convergence of AI and the IoT

The opening vignette illustrated to us how AI technologies when combined with IoT can provide solutions to complex problems. IoT collects a large amount of data from sensors and other "things." These data need to be processed for decision support. Later we will see how Microsoft's Cortana does this. Butner (2018) describes how combining AI and IoT can lead to the "next-level solutions and experiences." The emphasis in such combination is on learning more about customers and their needs. This integration also can facilitate competitive analysis and business operation (see the opening vignette). The combined pair of AI and IoT, especially when combined with Big Data, can help facilitate the discovery of new products, business processes, and opportunities. The full potential of IoT can be leveraged with AI technologies. In addition, the only way to make sense of the data streamed from the "things" via IoT and to obtain the insight from them is to subject them to AI analysis. Faggela (2017) provides the following three examples of combining AI and IoT:

1. The smart thermostat of Nest Labs (see smart homes in Chapter 13).
2. Automated vacuum cleaners, like iRobot Roomba (see Chapter 2, intelligent vacuums).
3. Self-driving vehicles (see Chapter 13).

The IoT can become very intelligent when combined with IBM Watson Analytics that includes machine learning. Examples are presented in the opening vignette and the opening vignette to Chapter 13.

The Convergence with Blockchain and Other Technologies

Several experts raise the possibility of the convergence of AI, analytics, and blockchain (e.g., Corea, 2017; Kranz, 2017). The idea is that such convergence may contribute to design or redesign of paradigms and technologies. The blockchain technology can add security to data shared by all parties in a distributed network, where transaction data can be recorded. Kranz believes that the conversion with blockchain will power new solutions to complex problems. Such a convergence should include the IoT. Kranz also see a role for fog computing (Chapter 9). Such a combination can be very useful in complex applications such as autonomous vehicles and in Amazon's Go (Application Case 1.10).

Application Case 1.10 Amazon Go Is Open for Business

In early 2018, **Amazon.com** opened its first fully automated convenience store in downtown Seattle. The company had had success with this type of store during 2017, experimenting with only the company's employees.

Shoppers enter the store, pick up products, and go home. Their accounts are charged later on. Sounds great! No more waiting in line for the packing of your goods and paying for them – no cashiers, no hassle.

In some sense, shoppers are going through a process similar to what they do online—find desired products/services, buy them, and wait for the monthly electronic charge.

The Shopping Process

To participate, you need a special free app on your smartphone. You need to connect it to your regular **Amazon.com** account. Here is what you do next:

1. Open your app.
2. Wave your smartphone at a gate to the store. It will work with a QR code there.
3. Enter the store.
4. Start shopping. All products are prepacked. You put them in a shopping bag (yours or one borrowed at the store). The minute you pick an item from the shelf, it is recorded in a virtual shopping cart. This activity is done by sensors/cameras. Your account is debited. If you change your mind, and return an item, the system will credit your account instantly. The sensors also track your movements in the store. (This is an issue of digital privacy; see Chapter 14, Section 14.3). The sensors are of RFID type (Chapter 13).
5. Finished shopping? Just leave the store (make sure your app is open for the gate to let you leave). The system knows that you have left and what products you took, and your shopping trip is finished. The system will total your cost, which you can check anytime on your smartphone.
6. **Amazon.com** records your shopping habits (again, a privacy issue), which will help your future shopping experience and will help Amazon to build recommendations for you (Chapter 2). The objective of Go is to guide you to healthy food! (Amazon sells its meal kits of healthy food there.)

Note: Today, only few people work in the store! Employees stock shelves and assist you otherwise. The company plans to open several additional stores in 2018.

The Technology Used

Amazon disclosed some of the technologies used. These are deep learning algorithms, computer vision, and sensor fusion. Other technologies were not disclosed. See the **videoyoutube.com/watch?v=NrmMk1Myrxc** (1:50 min.).

Sources: Condensed forC. Jarrett. (2018). "Amazon Set to Open Doors on AI-Powered Grocery Store." **Venturebeat.com**. **venturebeat.com/2018/01/21/amazon-set-to-open-doors-on-ai-powered-grocery-store/** (accessed September 2018); D. Reisinger. (2018, February 22). "Here Are the Next Cities to Get Amazon Go Cashier-Less Stores." *Fortune*.

Questions for Case 1.9

1. Watch the video. What did you like in it, and what did you dislike?
2. Compare the process described here to a self-check available today in many supermarkets and "big box" stores (Home Depot, etc.).
3. The store was opened in downtown Seattle. Why was the downtown location selected?
4. What are the benefits to customers? To Amazon?
5. Will customers be ready to trade privacy for convenience? Discuss.

For a comprehensive report regarding convergence of intelligent technologies, see **reportbuyer.com/product/5023639/**.

In addition to blockchain, one can include IoT and Big Data, as suggested earlier, as well as more intelligent technologies (e.g., machine vision, voice technologies). These may have enrichment effects. In general, the more technologies are used (presumably properly), the more complex problems may be solved, and the more efficient the performance of the convergence systems (e.g., speed, accuracy) will be. For a discussion, see **i-scoop.eu/convergence-ai-iot-big-data-analytics/**.

IBM and Microsoft Support for Intelligent Systems Convergence

Many companies provide tools or platforms for supporting intelligent systems convergence. Two examples follow.

IBM IBM is combining two of its platforms to support the convergence of AI and analytics. Power AI is a distribution platform for AI and machine learning. This is a way to support the IBM analytics platform called Data Science Experience (cloud enabled). The combination of the two enables improvements in data analytics process. It also enables data scientists to facilitate the training of complex AI models and neural networks. Researchers can use the combined system for deep learning projects. All in all, this combination provides better insight to problem solving. For details, see FinTech Futures (2017).

As you may recall from the opening vignette, IBM Watson is also combining analytics, AI, and IoT in cognitive buildings projects.

MICROSOFT'S CORTANA INTELLIGENCE SUITE Microsoft offers from its AZURE cloud (Chapter 13) a combination of advanced analytics, traditional BI, and Big Data analytics. The suite enables users to transform data into intelligent actions.

Using Cortana, one can transform data from several sources, including from IoT sensors, and apply both advanced analytics (e.g., data mining) and AI (e.g., machine learning) and extract insights and actionable recommendations, which are delivered to decision makers, to apps, or to fully automated systems. For the details of the system and the architecture of Cortana, see **mssqltips.com/sqlservertip/4360/introduction-to-microsoft-cortana-intelligence-suite/**.

SECTION 1.8 REVIEW QUESTIONS

1. What are the major benefits of intelligent systems convergences?
2. Why did analytics initiatives fail at such a high rate in the past?
3. What synergy can be created by combining AI and analytics?
4. Why is Big Data preparation essential for AI initiatives?
5. What are the benefits of adding IoT to intelligent technology applications?
6. Why it is recommended to use blockchain in support of intelligent applications?

1.9 OVERVIEW OF THE ANALYTICS ECOSYSTEM

So you are excited about the potential of analytics, data science, and AI and want to join this growing industry. Who are the current players, and what to do they do? Where might you fit in? The objective of this section is to identify various sectors of the analytics industry, provide a classification of different types of industry participants, and illustrate the types of opportunities that exist for analytics professionals. Eleven different types of players are identified in an **analytics ecosystem**. An understanding of the ecosystem also gives the reader a broader view of how the various players come together. A secondary

purpose of understanding the analytics ecosystem for a professional is also to be aware of organizations and new offerings and opportunities in sectors allied with analytics.

Although some researchers have distinguished business analytics professionals from data scientists (Davenport and Patil, 2012), as pointed out previously, for the purpose of understanding the overall analytics ecosystem, we treat them as one broad profession. Clearly, skill needs can vary for a strong mathematician to a programmer to a modeler to a communicator, and we believe this issue is resolved at a more micro/individual level rather than at a macro level of understanding the opportunity pool. We also take the widest definition of analytics to include all three types as defined by INFORMS—descriptive/reporting/visualization, predictive, and prescriptive as described earlier. We also include AI within this same pool.

Figure 1.17 illustrates one view of the analytics ecosystem. The components of the ecosystem are represented by the petals of an analytics flower. Eleven key sectors or clusters in the analytics space are identified. The components of the analytics ecosystem are grouped into three categories represented by the inner petals, outer petals, and the seed (middle part) of the flower. The outer six petals can be broadly termed *technology providers*. Their primary revenue comes from providing technology, solutions, and training to analytics user organizations so they can employ these technologies in the most effective and efficient manner. The inner petals can be generally defined as the *analytics accelerators*. The accelerators work with both technology providers and users. Finally, the core of the ecosystem comprises the *analytics user organizations*. This is the most important component as every analytics industry cluster is driven by the user organizations.

The metaphor of a flower is well suited for the analytics ecosystem as multiple components overlap each other. Similar to a living organism like a flower, all these petals grow and wither together. Many companies play in multiple sectors within the analytics industry and thus offer opportunities for movement within the field both horizontally and vertically.

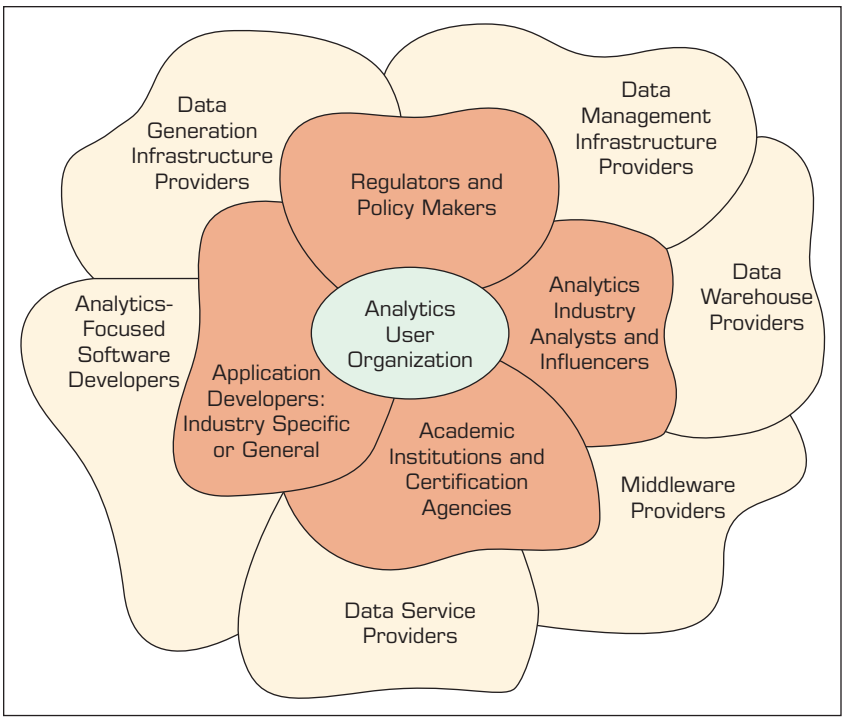


FIGURE 1.17 **Analytics Ecosystem.**

More details for the analytics ecosystem are included in our shorter book (Sharda, Delen, and Turban, 2017) as well as in Sharda and Kalgotra (2018). Matt Turck, a venture capitalist with FirstMark, has also developed and updates an analytics ecosystem focused on Big Data. His goal is to keep track of new and established players in various segments of the Big Data industry. A very nice visual image of his interpretation of the ecosystem and a comprehensive listing of companies is available through his Web site: **http://mattturck.com/2016/02/01/big-data-landscape/** (accessed September 2018).

1.10 PLAN OF THE BOOK

The previous sections have given you an understanding of the need for information technology in decision making, the evolution of BI, analytics, data science, and artificial intelligence. In the last several sections, we have seen an overview of various types of analytics and their applications. Now we are ready for a more detailed managerial excursion into these topics along with some deep hands-on experience in some of the technical topics. Figure 1.18 presents a plan on the rest of the book.

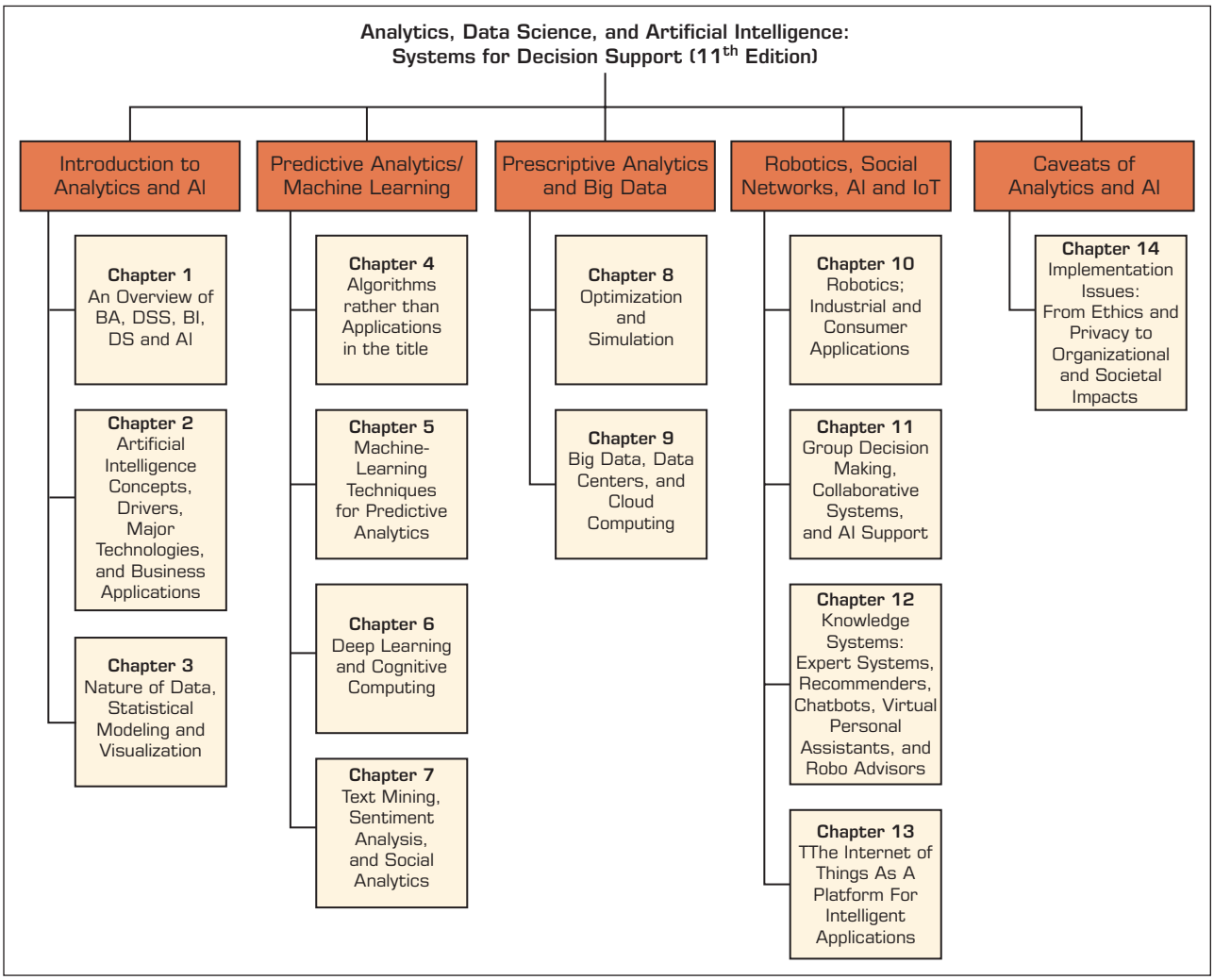


FIGURE 1.18 Plan of the Book.

In this chapter, we have provided an introduction, definitions, and overview of DSS, BI, and analytics, including Big Data analytics and data science. We also gave you an overview of the analytics ecosystem to have you appreciate the breadth and depth of the industry. Chapters 2 and 3 cover descriptive analytics and data issues. Data clearly form the foundation for any analytics application. Thus, we cover an introduction to data warehousing issues, applications, and technologies. This chapter also covers business reporting and visualization technologies and applications.

We follow the current chapter with a deeper introduction to artificial intelligence in Chapter 2. Because data are fundamental to any analysis, Chapter 3 introduces data issues as well as descriptive analytics, including statistical concepts and visualization. An online chapter covers data warehousing processes and fundamentals for those who like to dig more deeply into these issues. The next section of the book covers predictive analytics and machine learning. Chapter 4 provides an introduction to data mining applications and the data mining process. Chapter 5 introduces many of the common data mining techniques: classification, clustering, association mining, and so forth. Chapter 6 includes coverage of deep learning and cognitive computing. Chapter 7 focuses on text mining applications as well as Web analytics, including social media analytics, sentiment analysis, and other related topics. The following section brings the "data science" angle into further depth. Chapter 8 covers prescriptive analytics. Chapter 9 includes more details of Big Data analytics. It also includes an introduction to cloud-based analytics as well as location analytics. The next section covers robotics, social networks, AI, and IoT. Chapter 10 introduces robots in business and consumer applications and discusses the future impact of such devices on society. Chapter 11 focuses on collaboration systems, crowdsourcing, and social networks. Chapter 12 reviews personal assistants, chatbots, and the exciting developments in this space. Chapter 13 studies IoT and its potential in decision support and a smarter society. The ubiquity of wireless and GPS devices and other sensors is resulting in the creation of massive new databases and unique applications. A new breed of analytics companies is emerging to analyze these new databases and create a much better and deeper understanding of customers' behaviors and movements. It is leading to the automation of analytics and has spanned a new area called the "Internet of Things." Finally, Chapter 14 concludes with a brief discussion of security, privacy, and societal dimensions of analytics/AI.

1.11 RESOURCES, LINKS, AND THE TERADATA UNIVERSITY NETWORK CONNECTION

The use of this chapter and most other chapters in this book can be enhanced by the tools described in the following sections.

Resources and Links

We recommend the following major organizational resources and links:

- The Data Warehousing Institute (**tdwi.org**).
- Data Science Central (**datasciencecentral.com**).
- DSS Resources (**dssresources.com**).
- Microsoft Enterprise Consortium (**enterprise.waltoncollege.uark.edu/mec.asp**).

Vendors, Products, and Demos

Most vendors provide software demos of their products and applications. Information about products, architecture, and software is available at **dssresources.com**.

Periodicals

We recommend the following periodicals:

- *Decision Support Systems* (**www.journals.elsevier.com/decision-support-systems**).
- *CIO Insight* (**www.cioinsight.com**).

The Teradata University Network Connection

This book is tightly connected with the free resources provided by TUN (see **www.teradatauniversitynetwork.com**). The TUN portal is divided into two major parts: one for students and one for faculty. This book is connected to the TUN portal via a special section at the end of each chapter. That section includes appropriate links for the specific chapter, pointing to relevant resources. In addition, we provide hands-on exercises using software and other materials (e.g., cases) available at TUN.

The Book's Web Site

This book's Web site, pearsonhighered.com/sharda, contains supplemental textual material organized as Web chapters that correspond to the printed book's chapters. The topics of these chapters are listed in the online chapter table of contents.

As this book went to press, we verified that all cited Web sites were active and valid. However, URLs are dynamic. Web sites to which we refer in the text sometimes change or are discontinued because companies change names, are bought or sold, merge, or fail. Sometimes Web sites are down for maintenance, repair, or redesign. Many organizations have dropped the initial "www" designation for their sites, but some still use it. If you have a problem connecting to a Web site that we mention, please be patient and simply run a Web search to try to identify a possible new site. Most times, you can quickly find the new site through one of the popular search engines. We apologize in advance for this inconvenience.

Chapter Highlights

- The business environment is becoming more complex and is rapidly changing, making decision making more difficult.
- Businesses must respond and adapt to the changing environment rapidly by making faster and better decisions.
- A model is a simplified representation or abstraction of reality.
- Decision making involves four major phases: intelligence, design, choice, and implementation.
- In the intelligence phase, the problem (opportunity) is identified, classified, and decomposed (if needed), and problem ownership is established.
- In the design phase, a model of the system is built, criteria for selection are agreed on, alternatives are generated, results are predicted, and a decision methodology is created.
- In the choice phase, alternatives are compared, and a search for the best (or a good-enough) solution is launched. Many search techniques are available.
- In implementing alternatives, a decision maker should consider multiple goals and sensitivity-analysis issues.
- The time frame for making decisions is shrinking, whereas the global nature of decision making is expanding, necessitating the development and use of computerized DSS.
- An early decision support framework divides decision situations into nine categories, depending on the degree of structuredness and managerial activities. Each category is supported differently.
- Structured repetitive decisions are supported by standard quantitative analysis methods, such as MS, MIS, and rule-based automated decision support.
- DSS use data, models, and sometimes knowledge management to find solutions for semistructured and some unstructured problems.

- The major components of a DSS are a database and its management, a model base and its management, and a user-friendly interface. An intelligent (knowledge-based) component can also be included. The user is also considered to be a component of a DSS.
- BI methods utilize a central repository called a DW that enables efficient data mining, OLAP, BPM, and data visualization.
- BI architecture includes a DW, business analytics tools used by end users, and a user interface (such as a dashboard).
- Many organizations employ descriptive analytics to replace their traditional flat reporting with interactive reporting that provides insights, trends, and patterns in the transactional data.
- Predictive analytics enables organizations to establish predictive rules that drive the business outcomes through historical data analysis of the existing behavior of the customers.
- Prescriptive analytics helps in building models that involve forecasting and optimization techniques based on the principles of OR and management science to help organizations to make better decisions.
- Big Data analytics focuses on unstructured, large data sets that may also include vastly different types of data for analysis.
- Analytics as a field is also known by industry-specific application names, such as sports analytics. It is also known by other related names such as data science or network science.
- Healthcare and retail chains are two areas where analytics applications abound, with much more to come.
- Image analytics is a rapidly evolving field leading to many applications of deep learning.
- The analytics ecosystem can be first viewed as a collection of providers, users, and facilitators. It can be broken into 11 clusters.

Key Terms

analytics
analytics ecosystem
artificial intelligence
augmented intelligence
Big Data analytics
business intelligence (BI)
choice phase
dashboard
data mining
decision or normative analytics
descriptive (or reporting) analytics
design phase
implementation phase
intelligence phase
online analytical processing (OLAP)
online transaction processing (OLTP)
predictive analytics
prescriptive analytics

Questions for Discussion

1. Survey the literature from the past six months to find one application each for DSS, BI, and analytics. Summarize the applications on one page, and submit it with the exact sources.
2. Your company is considering opening a branch in China. List typical activities in each phase of the decision (intelligence, design, choice, and implementation) regarding whether to open a branch.
3. You are about to buy a car. Using Simon's (1977) four-phase model, describe your activities at each step in making the decision.
4. Explain, through an example, the support given to decision makers by computers in each phase of the decision process.
5. Comment on Simon's (1977) philosophy that managerial decision making is synonymous with the whole process of management. Does this make sense? Explain. Use a real-world example in your explanation.
6. Review the major characteristics and capabilities of DSS. How does each of them relate to the major components of DSS?
7. List some internal data and external data that could be found in a DSS for a university's admissions office.
8. Distinguish BI from DSS.
9. Compare and contrast predictive analytics with prescriptive and descriptive analytics. Use examples.
10. Discuss the major issues in implementing BI.

Exercises

Teradata University Network and Other Hands-On Exercises

1. Go to the TUN site **teradatauniversitynetwork.com**. Using the site password your instructor provides, register for the site if you have not already previously registered. Log on and learn the content of the site. You will receive assignments related to this site. Prepare a list of 20 items on the site that you think could be beneficial to you.
2. Go to. Explore the Sports Analytics page, and summarize at least two applications of analytics in any sport of your choice.
3. Go to. The TUN site, and select "Cases, Projects, and Assignments." Then select the case study "Harrah's High Payoff from Customer Information." Answer the following questions about this case:
   a. What information does the data mining generate?
   b. How is this information helpful to management in decision making? (Be specific.)
   c. List the types of data that are mined.
   d. Is this a DSS or BI application? Why?
4. Go to **teradatauniversitynetwork.com** and find the paper titled "Data Warehousing Supports Corporate Strategy at First American Corporation" (by Watson, Wixom, and Goodhue). Read the paper, and answer the following questions:
   a. What were the drivers for the DW/BI project in the company?
   b. What strategic advantages were realized?
   c. What operational and tactical advantages were achieved?
   d. What were the critical success factors for the implementation?
5. Go to **http://analytics-magazine.org/issues/digital-editions** and find the January/February 2012 edition titled "Special Issue: The Future of Healthcare." Read the article "Predictive Analytics—Saving Lives and Lowering Medical Bills." Answer the following questions:
   a. What problem is being addressed by applying predictive analytics?
   b. What is the FICO Medication Adherence Score?
   c. How is a prediction model trained to predict the FICO Medication Adherence Score HoH? Did the prediction model classify the FICO Medication Adherence Score?
   d. Zoom in on Figure 4, and explain what technique is applied to the generated results.
   e. List some of the actionable decisions that were based on the prediction results.
6. Go to **http://analytics-magazine.org/issues/digital-editions**, and find the January/February 2013 edition titled "Work Social." Read the article "Big Data, Analytics and Elections," and answer the following questions:
   a. What kinds of Big Data were analyzed in the article's Coo? Comment on some of the sources of Big Data.
   b. Explain the term *integrated system*. What is the other technical term that suits an *integrated system*?
   c. What data analysis techniques are employed in the project? Comment on some initiatives that resulted from data analysis.
   d. What are the different prediction problems answered by the models?
   e. List some of the actionable decisions taken that were based on the prediction results.
   f. Identify two applications of Big Data analytics that are not listed in the article.
7. Search the Internet for material regarding the work of managers and the role analytics plays in it. What kinds of references to consulting firms, academic departments, and programs do you find? What major areas are represented? Select five sites that cover one area, and report your findings.
8. Explore the public areas of **dssresources.com**. Prepare a list of its major available resources. You might want to refer to this site as you work through the book.
9. Go to **microstrategy.com**. Find information on the five styles of BI. Prepare a summary table for each style.
10. Go to **oracle.com**, and click the Hyperion link under Applications. Determine what the company's major products are. Relate these to the support technologies cited in this chapter.
11. Go to the TUN questions site. Look for BSI videos. Review the video of "Case of Retail Tweeters." Prepare a one-page summary of the problem, proposed solution, and the reported results. You can also find associated slides on **slideshare.net**.
12. Review the Analytics Ecosystem section. Identify at least two additional companies in at least five of the industry clusters noted in the discussion.
13. The discussion for the analytics ecosystem also included several typical job titles for graduates of analytics and data science programs. Research Web sites such as **datasciencecentral.com** and **tdwi.org** to locate at least three similar job titles that you may find interesting for your career.
14. Go to Brainspace at MIT lab **brainspace.com**. View the video about "Augmented Human Intelligence." Find the activities that deal with the enabling of meaningful combination of people and machines. Write a report.
15. Find information about IBM Watson's activities in the healthcare field. Write a report.
16. Examine Daniel Power's DSS Resources site at **dssresources.com**. Take the Decision Support Systems Web Tour (**dssresources.com/tour/index.html**). Explore other areas of the Web site. List at least three recent resources related to analytics. What topics do these cover?

References

http://canopeoapp.com/ (accessed October 2018).

http://imazon.org.br/en/imprensa/mapping-change-in-the-amazon-how-satellite-images-are-halting-deforestation/ (accessed October 2018).

http://www.earthcastdemo.com/2018/07/bloomberg-earthcast-customizing-weather/ (accessed October 2018)

https://www.worldbank.org/en/news/press-release/2018/02/22/world-bank-supports-sierra-leones-efforts-in-landslide-recovery (accessed October 2018)

Siemens.com. **About Siemens. siemens.com/about/en/** (accessed September 2018).

Silvaris.com. **Silvaris overview. silvaris.com** (accessed September 2018).

Agah, A. (2017). *Medical Applications of Artificial Intelligence*. Boca Raton, FL: CRC Press.

Anthony, R. N. (1965). *Planning and Control Systems: A Framework for Analysis*. Cambridge, MA: Harvard University Graduate School of Business.

Baker, J., and M. Cameron. (1996, September). "The Effects of the Service Environment on Affect and Consumer Perception of Waiting Time: An Integrative Review and Research Propositions." *Journal of the Academy of Marketing Science, 24*, pp. 338–349.

Bean, R. (2017, May 8). "How Big Data Is Empowering AI and Machine Learning at Scale." *MIT Sloan Management Review*.

Boddington, P. (2017, January 3). "Case Study: Robot Camel Jockeys. Yes, really." *Ethics for Artificial Intelligence*.

Brainspace. (2016, June 13). "Augmenting Human Intelligence." *MIT Technology Review Insights*.

Butner, K. (2018, January 8). "Combining Artificial Intelligence with the Internet of Things Could Make Your Business Smarter." IBM Consulting Blog.

CDC.gov. (2015, September 21). "Important Facts about Falls." **cdc.gov/homeandrecreationalsafety/falls/adultfalls.html** (accessed September 2018).

Charles, T. (2018, May 21). "Influence of the External Environment on Strategic Decision." *Azcentral*. **yourbusiness.azcentral.com/influence-external-environment-strategic-decisions-17628.html/** (accessed October 2018).

Chiguluri, V., Guthikonda, K., Slabaugh, S., Havens, E., Peña, J., & Cordier, T. (2015, June). Relationship Between Diabetes Complications and Health Related Quality of Life Among an Elderly Population in the United States. Poster presentation at the American Diabetes Association Seventy-Fifth Annual Scientific Sessions. Boston, MA.

Chongwatpol, J., & R. Sharda. (2010, December). "SNAP: A DSS to Analyze Network Service Pricing for State Networks." *Decision Support Systems, 50*(1), pp. 347–359.

Chung, C. (2016). "Dubai Camel Race Ride-Along." **YouTube.com**. **youtube.com/watch?v=xFCRhk4GYds** (accessed September 2018).

Cordier, T., Slabaugh, L., Haugh, G., Gopal, V., Cusano, D., Andrews, G., & Renda, A. (2015, September). Quality of Life Changes with Progressing Congestive Heart Failure. Poster presentation at the Nineteenth Annual Scientific Meeting of the Heart Failure Society of America, Washington, DC.

Corea, F. (2017, December 1). "The Convergence of AI and Blockchain: What's the Deal?" **Medium.com**. **medium.com/@Francesco_AI/the-convergence-of-ai-and-blockchain-whats-the-deal-60c618e3accc** (accessed September 2018).

Davenport, T., & SAS Institute Inc. (2014, February). "Analytics in Sports: The New Science of Winning." **sas.com/content/dam/SAS/en_us/doc/whitepaper2/iia-analytics-in-sports-106993.pdf** (accessed September 2018).

Davenport, T. H., & Patil, D. J. (2012). "Data Scientist." *Harvard Business Review, 90*, 70–76.

Deahl, D. (2018, January 7). "This Automatic Feeder Can Tell the Difference Between Your Pets." *The Verge*.

De Smet, A., et al. (2017, June). "Untangling Your Organization's Decision Making." *McKinsey Quarterly*.

Duncan, A. (2016). "The BICC Is Dead." **https://blogs.gartner.com/alan-duncan/2016/03/11/the-bicc-is-dead/** (accessed October 2018).

Dundas.com. "How Siemens Drastically Reduced Cost with Managed BI Applications." **www.dundas.com/Content/pdf/siemens-case-study.pdf** (accessed September 2018).

Eckerson, W. (2003). *Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions*. Seattle, WA: The Data Warehousing Institute.

eMarketer. (2017, May). "Artificial Intelligence: What's Now, What's New and What's Next?" EMarketer Inc.

Emc.com. (n.d.). "Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field." **emc.com/collateral/about/news/emc-data-science-study-wp.pdf** (accessed September 2018)

Emmert, Samantha. (2018, March 19). "Fighting Illegal Fishing." Global Fishing Watch. **globalfishingwatch.org/research/fighting-illegal-fishing/** (accessed October 2018).

Faggela, D. (2017, August 24). "Artificial Intelligence Plus the Internet of Things (IoT): 3 Examples Worth Learning From." *TechEmergence*.

Faggela, D. (2018, March 29). "Artificial Intelligence Industry: An Overview by Segment." *TechEmergence*.

Fernandez, J. (2017, April). "A Billion People a Day. Millions of Elevators. No Room for Downtime." IBM developer Works Blog. **developer.ibm.com/dwblog/2017/kone-watson-video/** (accessed September 2018).

FinTech Futures. (2017, October 11). "IBM Combining Data Science and AI for Analytics Advance." **BankingTech.com**.

Gartner, Inc. (2004). Using Business Intelligence to Gain a Competitive Edge. A special report.

Gates, S., Smith, L. A., Fisher, J. D., et al. (2008). Systematic Review of Accuracy of Screening Instruments for Predicting Fall Risk Among Independently Living Older Adults. *Journal of Rehabilitation Research and Development, 45*(8), pp. 1105–1116.

Gill, T. M., Murphy, T. E., Gahbauer, E. A., et al. (2013). "Association of Injurious Falls with Disability Outcomes and Nursing Home Admissions in Community Living Older Persons." *American Journal of Epidemiology, 178*(3), pp. 418–425.

Gorry, G. A., & Scott-Morton, M. S. (1971). "A Framework for Management Information Systems." *Sloan Management Review, 13*(1), pp. 55–70.

Havens, E., Peña, J., Slabaugh, S., Cordier, T., Renda, A., & Gopal, V. (2015, October). Exploring the Relationship Between Health-Related Quality of Life and Health Conditions, Costs, Resource Utilization, and Quality Measures. Podium presentation at the ISOQOL Twenty-Seventh Annual Conference, Vancouver, Canada.

Havens, E., Slabaugh, L., Peña, J., Haugh, G., & Gopal, V. (2015, February). Are There Differences in Healthy Days Based on Compliance to Preventive Health Screening Measures? Poster presentation at Preventive Medicine 2015, Atlanta, GA.

Healthcare IT News. (2017, November 9). "How AI Is Transforming Healthcare and Solving Problems in 2017." Slideshow. **healthcareitnews.com/slideshow/how-ai-transforming-healthcare-and-solving-problems-2017?page=4/** (accessed September 2018).

Hesse, R., & G. Woolsey. (1975). *Applied Management Science: A Quick and Dirty Approach*. Chicago, IL: SRA Inc.

Humana. 2016 Progress Report. **populationhealth.humana.com/wp-content/uploads/2016/05/BoldGoal2016ProgressReport_1.pdf** (accessed September 2018).

INFORMS. Analytics Section Overview. **informs.org/Community/Analytics** (accessed September 2018).

Jacquet, F. (2017, July 4). "Exploring the Artificial Intelligence Ecosystem: AI, Machine Learning, and Deep Learning." DZone.

Jarrett, C. (2018, January 21). "Amazon Set to Open Doors on AI-Powered Grocery Store." **Venturebeat.com**. **venturebeat.com/2018/01/21/amazon-set-to-open-doors-on-ai-powered-grocery-store/** (accessed September 2018).

Keen, P. G. W., & M. S. Scott-Morton. (1978). Decision Support Systems: An Organizational Perspective. Reading, MA: Addison-Wesley.

Kemper, C., and C. Breuer. (2016). "How Efficient Is Dynamic Pricing for Sports Events? Designing a Dynamic Pricing Model for Bayern Munich." *International Journal of Sports Finance,* 11, pp. 4–25.

Kranz, M. (2017, December 27). "In 2018, Get Ready for the Convergence of IoT, AI, Fog, and Blockchain." *Insights*.

Liberto, D. (2017, June 29). "Artificial Intelligence Will Add Trillion to the Global Economy: PwC." *Investopedia*.

Lollato, R., Patrignani, A., Ochsner, T. E., Rocatelli, A., Tomlinson, P. & Edwards, J. T. (2015). "Improving Grazing Management Using a Smartphone App." **www.bookstore.ksre.ksu.edu/pubs/MF3304.pdf** (accessed October 2018).

Lopez, J. (2017, August 11). "Smart Farm Equipment Helps Feed the World." **IQintel.com**.

Metz, C. (2018, February 12). "As China Marches Forward on A.I., the White House Is Silent." *The New York Times*.

Nadav, S. (2017, August 9). "Business Intelligence Is Failing; Here Is What Is Coming Next." *Huffington Post*.

Norman, A. (2018, January 31). "Your Future Doctor May Not Be Human. This Is the Rise of AI in Medicine." **Futurism.com**.

Olshansky, C. (2017, August 24). "Coca-Cola Is Bringing Artificial Intelligence to Vending Machines." *Food & Wine*.

Opfer, C. (2016, June 22). "There's One Terrific Reason to Race Camels Using Robot Jockeys." **Howstuffworks.com**.

Padmanabhan, G. (2018, January 4). "Industry-Specific Augmented Intelligence: A Catalyst for AI in the Enterprise." *Forbes*.

Pajouh Foad, M., Xing, D., Hariharan, S., Zhou, Y., Balasundaram, B., Liu, T., & Sharda, R. (2013). Available-to-Promise in Practice: An Application of Analytics in the Specialty Steel Bar Products Industry. *Interfaces, 43*(6), pp. 503–517. **dx.doi.org/10.1287/inte.2013.0693** (accessed September 2018).

Patrignani, A., & Ochsner, T. E., (2015). Canopeo: A Powerful New Tool for Measuring Fractional Green Canopy Cover. *Agronomy Journal, 107*(6), pp. 2312–2320;

PE Report. (2017, July 29). "Satellite-Based Advance Can Help Raise Farm Output by 20 Percent Experts." *Financial Express*.

PricewaterhouseCoopers Report. (2011, December). "Changing the Game: Outlook for the Global Sports Market to 2015." **pwc.com/gx/en/hospitality-leisure/pdf/changing-the-game-outlook-for-the-global-sports-market-to-2015.pdf** (accessed September 2018).

Quain, S. (2018, June 29). "The Decision-Making Process in an Organization." *Small Business Chron*.

Reisinger, D. (2018, February 22). "Here Are the Next Cities to Get Amazon Go Cashier-Less Stores." *Fortune*.

Sharda, R., Asamoah, D., & Ponna, N. (2013). "Research and Pedagogy in Business Analytics: Opportunities and Illustrative Examples." *Journal of Computing and Information Technology, 21*(3), pp. 171–182.

Sharda, R., Delen, D., & Turban, E. (2016). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective on Analytics*. 4th ed. NJ: Pearson.

Sharda, R., & P. Kalgotra. (2018). "The Blossoming Analytics Talent Pool: An Overview of the Analytics Ecosystem." In James J. Cochran (ed.). *INFORMS Analytics Body of Knowledge. John Wiley, Hoboken, NJ*

Silk, R. (2017, November). "Biometrics: Facial Recognition Tech Coming to an Airport Near You." *Travel Weekly, 21*.

Simon, H. (1977). *The New Science of Management Decision*. Englewood Cliffs, NJ: Prentice Hall.

Slade, L. (2017, December 21). "Meet the Jordanian Camel Races Using Robot Jockeys." **Sbs.com**.au.

Slowey, L. (2017, February 16). "Look Who's Talking: KONE Makes Elevator Services Truly Intelligent with Watson IoT." IBM Internet of Things Blog. **ibm.com/blogs/internet-of-things/kone/** (accessed September 2018).

"Sports Analytics Market Worth by 2021." (2015, June 25). Wintergreen Research Press Release. Covered by PR Newswire at **http://www.prnewswire.com/news-releases/sports-analytics-market-worth-47-billion-by-2021-509869871.html**.

Srikanthan, H. . (2018, January 8). "KONE Improves 'People Flow' in 1.1 Million Elevators with IBM Watson IoT." Generis. **https://generisgp.com/2018/01/08/ibm-case-study-kone-corp/** (accessed September 2018).

Staff. "Assisted, Augmented and Autonomous: The 3Flavours of AI Decisions." (2017, June 28). *Software and Technology*.

Tableau.com. Silvaris Augments Proprietary Technology Platform with Tableau's Real-Time Reporting Capabilities. **tableau.com/sites/default/files/case-studies/silvaris-business-dashboards_0.pdf** (accessed September 2018).

Tartar, Andre, et al. (2018, 26 July). "All the Things Satellites Can Now See from Space." **Bloomberg.com**. **www.bloomberg.com/news/features/2018-07-26/all-the-things-satellites-can-now-see-from-space** (accessed October 2018).

TeradataUniversityNetwork.com. (2015, Fall). "BSI: Sports Analytics—Precision Football" (video). teradatauniversity **network.com/About-Us/Whats-New/BSI–Sports-Analytics-Precision-Football/** (accessed September 2018).

Thibodeaux, W. (2017, June 29). "This Artificial Intelligence Kiosk Is Designed to Spot Liars at Airports." **Inc.com**.

Turck, Matt. "Is Big Data Still a Thing? (The 2016 Big Data Landscape)." **http://mattturck.com/2016/02/01/big-data-landscape/** (accessed September 2018).

Watson, H. (2005, Winter). Sorting Out What's New in Decision Support. *Business Intelligence Journal*.

Weldon, D. (2018, March 6). "Nearly Half of CIOs Now Plan to Deploy Artificial Intelligence." Information Management.

Wikipedia.org. On-base Percentage. **wikipedia.org/wiki/On_base_percentage**. (accessed September 2018).

Wikipedia.org. Sabermetrics. **wikipedia.org/wiki/Sabermetrics** (accessed September 2018).

Wikipedia.org. SIEMENS. **wikipedia.org/wiki/Siemens** (accessed September 2018).

YouTube.com. (2013, December 17). CenterPoint Energy Talks Real Time Big Data Analytics. **youtube.com/watch?v=s7CzeSlIEfI** (accessed September 2018).

Yue, P. (2017, August 24). "Baidu, Beijing Airport Launch Facial Recognition for Passenger Check-In." China Money Network. **https://www.chinamoneynetwork.com/2017/08/24/baidu-capital-airport-launch-facial-recognition-system-airport** (accessed October 2018).

Zane, E. B. (2016). *Effective Decision-Making: How to Make Better Decisions Under Uncertainty And Pressure*. Kindle ed. Seattle, WA: Amazon Digital Services.

CHAPTER

2

Artificial Intelligence

Concepts, Drivers, Major Technologies, and Business Applications

LEARNING OBJECTIVES

- Understand the concepts of artificial intelligence (AI)
- Become familiar with the drivers, capabilities, and benefits of AI
- Describe human and machine intelligence
- Describe the major AI technologies and some derivatives
- Discuss the manner in which AI supports decision making
- Describe AI applications in accounting
- Describe AI applications in banking and financial services
- Describe AI in human resource management
- Describe AI in marketing
- Describe AI in production-operation management

Artificial intelligence (AI), which was a curiosity for generations, is rapidly developing into a major applied technology with many applications in a variety of fields. OpenAI's (an AI research institution described in Chapter 14) mission states that AI will be the most significant technology ever created by humans. AI appears in several shapes and has several definitions. In a crude way, it can be said that AI's aim is to make machines exhibit intelligence as close as possible to what people exhibit, hopefully for the benefit of humans. The latest developments in computing technologies drive AI to new levels and achievements. For example, *IDC Spending Guide* (March 22, 2018) forecasted that worldwide spending on AI will reach $19.1 billion in 2018. It also predicted annual double-digit spending growth for the near future. According to Sharma (2017), China expects to be the world leader in AI, with a spending of $60 billion in 2025. For the business value of AI, see Greig (2018).

In this chapter, we provide the essentials of AI, its major technologies, its support for decision making, and a sample of its applications in the major business functional areas.

The chapter has the following sections:

2.1 OPENING VIGNETTE: INRIX Solves Transportation Problems

THE PROBLEM

Traffic congestion is an ever-increasing problem in many large metropolitan areas. Drivers may spend several hours on the roads each day. In addition, air pollution is increasing, and more accidents are occurring.

THE SOLUTION

INRIX corporation (**inrix.com**) enables drivers to get real-time traffic information. They can download the INRIX-XD Traffic App for iOS and Android. The information provided is generated by a predictive analysis of massive data obtained from consumers and the environment (e.g., road construction, accidents). Information sources include:

- Traffic data collected by helicopters, drones, and so on, which include real-time traffic flow and accident information.
- Information provided by participating delivery companies and over 100 million anonymous volunteer drivers, who have GPS-enabled smartphones, all reporting in real time.
- Information provided by traffic congestion reports (e.g., delays due to road maintenance).

INRIX processes the collected information with proprietary analytical tools and formulas, some of which are AI-based. The processed information is used to generate traffic predictions. For example, it creates a picture of anticipated traffic flows and delays for the next 15 to 20 minutes, few hours, and few days for many locations. These predictions enable drivers to plan their optimal routes. As of 2018, INRIX had offered global coverage in 45 countries and in many major cities, and the company analyzed traffic information from over 100 sources. This service is combined with digital maps. In Seattle, for example, traffic information is disseminated via smartphones and color codes on billboards along the freeways. Smartphones also display estimated times for the roads to be either clear or jammed. As of 2018, the company had covered over 5,000,000 miles of highways worldwide, delivering upon request the best recommended routes to use, all in real time.

The INRIX system provides information (or recommendations) for decisions such as:

- Optional routes for delivery vehicles and other travelers to take
- The best time to go to work or to other places from a given location

- Information for rerouting a trip to avoid encountering a traffic jam that just occurred
- Fees to be paid on highways, which are based on traffic conditions and time of the day

The technologies used to collect data are:

- Closed-circuit TV cameras and radar that monitor traffic conditions
- Public safety reports and traffic information
- Information about freeway access and departure flows
- Technologies that measure toll collection queues
- Magnetic sensing detectors embedded under the road surface (expensive)
- Smartphones and other data collection devices that gather data for INRIX

The information is processed by several AI techniques such as expert systems; see Chapter 12 and different analytical models (such as simulation).

Several of the sources of information are connected to the company via the Internet of Things (IoT) (Chapter 13). According to its Web site, INRIX has partnered with Clear Channel Radio to broadcast real-time traffic data directly to vehicles via Ln Carr or via portable navigation systems, broadcast media, and wireless and Internet-based services. Clear Channel's Total Traffic Network is available in more than 125 metropolitan areas in four countries (**inrix.com/press-releases/2654/**). In 2018, the system was installed in over 275 million cars and data collection devices. The system collects real-time traffic information from these devices.

THE RESULTS

In addition to being used by individual drivers, the processed information is shared by organizations and city planners for making planning decisions. Also, less traffic congestion has been recorded in participating cities, which results in less pollution, fewer road accidents, and increased productivity by happier employees who spend less time commuting.

The INRIX Traffic App (available for download at **inrix.com/mobile-apps**) is suitable for all smartphones; it supports 10 languages, including English, French, and Spanish. For the free INRIX traffic features, see **inrixtraffic.com/features**. For interesting case studies, see **inrix.com/case-studies**.

As of 2016, INRIX had released an improved traffic app that uses both AI and crowdsourcing (Chapter 11) to support drivers' decisions as to the best route to take (Korosec, 2016). The AI technology analyzes drivers' historical activities to infer their future activities.

Note: Popular smartphone apps, such as Waze and Moovit, provide navigation and data collection similar to INRIX.

Sources: Based on **inrix.com**, Gitlin (2016), Korosec (2016), and **inrix.com/mobile-apps** (all accessed June 2018).

QUESTIONS FOR THE OPENING VIGNETTE

1. Explain why traffic may be down while congestion is up (see the London case at **inrix.com/uk-highways-agency/**).
2. How does this case relate to decision support?
3. Identify the AI elements in this system.
4. Identify developments related to AI by viewing the company's press releases from the most recent four months at **inrix.com/press-releases**. Write a report.

5. According to Gitlin (2016), INRIX's new mobile traffic app is a threat to Waze. Explain why.
6. Go to **sitezeus.com/data/inrix** and describe the relationship between INRIX and Zeus. View the 2:07 min. video at **sitezeus.com/data/inrix/**. Why is the system in the video called a "decision helper"?

WHAT WE CAN LEARN FROM THE VIGNETTE

The INRIX case illustrates to us how the collection and analysis of a very large amount of information (Big Data) can improve vehicles' mobility in large cities. Specifically, by collecting information from drivers and other sources instead of only from expensive sensors, INRIX has been able to optimize mobility. This has been achieved by supporting decisions made by drivers and by analyzing traffic flows. INRIX is also using applications from the IoT to connect vehicles and devices with its computing system. This application is one of the building blocks of smart cities (see Chapter 13). The analysis of the collected data is done by using powerful algorithms, some of which are applications of AI.

2.2 INTRODUCTION TO ARTIFICIAL INTELLIGENCE

We would all like to see computerized decision making being simpler, easier to use, more intuitive, and less threatening. And indeed, efforts have been made over time to simplify and automate several tasks in the decision-making process. Just think of the day that refrigerators will be able to measure and evaluate their contents and place orders for goods that need replenishment. Such a day is not too far in the future, and the task will be supported by AI.

CIO Insight projected that by 2035, intelligent computer technologies will result in $5–$8.3 trillion in economic value (see **cioinsight.com/blogs/how-ai-will-impact-the-global-economy.html**). Among the technologies listed as intelligent ones are the IoT, advanced robotics, and self-driven vehicles, all described in this book. Gartner, a leading technology consulting firm, listed the following in its 2016 and 2017 Hype Cycles for Emerging Technologies: expert advisors, natural language questions and answering, commercial drones, smart workspaces, IoT platforms, smart data discovery, general-purpose machine intelligence, and virtual personal assistants. Most are described or cited in this book (see also Greengard, 2016). For the history of AI, see Zarkadakis (2016) and **en.wikipedia.org/wiki/History_of_artificial_intelligence**.

Definitions

Artificial intelligence has several definitions (for an overview see Marr 2018); however, many experts agree that AI is concerned with two basic ideas: (1) the study of human thought processes (to understand what intelligence is) and (2) the representation and duplication of those thought processes in machines (e.g., computers, robots). That is, the machines are expected to have humanlike thought processes.

One well-publicized definition of AI is "the capabilities of a machine to imitate intelligent human behavior" (per *Merriam-Webster Dictionary*). The theoretical background of AI is based on logic, which is also used in several computer science innovations. Therefore, AI is considered a subfield of computer science. For the relationship between AI and logic, see **plato.stanford.edu/entries/logic-ai**.

A well-known early application of artificial intelligence was the chess program hosted at IBM's supercomputer (Deep Blue). The system beat the famous world champion, Grand Master Garry Kasparov.

AI is an umbrella term for many techniques that share similar capabilities and characteristics. For a list of 50 unique AI technologies, see Steffi (2017). For 33 types of AI, see **simplicable.com/new/types-of-artificial-intelligence**.

Major Characteristics of AI Machines

There is an increasing trend to make computers "smarter." For example, Web 3.0 supposes to enable computerized systems that exhibit significantly more intelligence than Web 2.0. Several applications are already based on multiple AI techniques. For example, the area of machine translation of languages is helping people who speak different languages to collaborate as well as to buy online products that are advertised in languages they do not speak. Similarly, machine translation can help people who know only their own language to converse with people speaking other languages and to make decisions jointly in real time.

Major Elements of AI

As described in Chapter 1, the landscape of AI is huge, including hundreds or more components. We illustrate the foundation and the major technologies in Figure 2.1. Notice that we divide them into two groups: Foundations, and Technologies and Applications. The major technologies will be defined later in this chapter and described throughout this book.

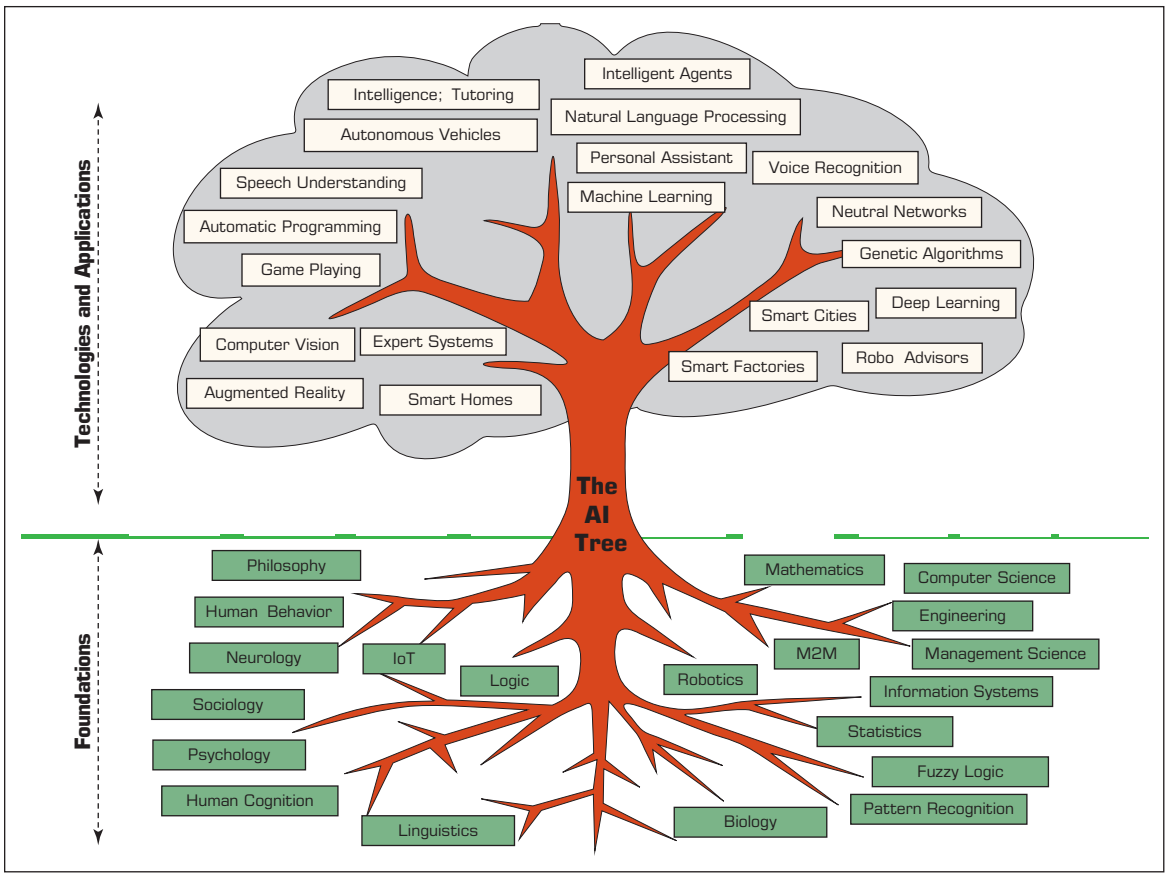


FIGURE 2.1 The Functionalities and Applications of Artificial Intelligence.

AI Applications

The technologies of AI are used in the creation of a large number of applications. In Sections 2.6–2.10, we provide a sampler of applications in the major functional areas of business.

Example

Smart or intelligent applications include those that can help machines to answer customers' questions asked in natural languages. Another area is that of knowledge-based systems which can provide advice, assist people to make decisions, and even make decisions on their own. For example, such systems can approve or reject buyers' requests to purchase online (if the buyers are not preapproved or do not have an open line of credit). Other examples include the automatic generating of online purchasing orders and arranging fulfillment of orders placed online. Both Google and Facebook are experimenting with projects that attempt to teach machines how to learn and support or even make autonomous decisions. For smart applications in enterprises, see Dodge (2016), Finlay (2017), McPherson (2017), and Reinharz (2017). For how AI solutions are used to facilitate government services, see BrandStudio (2017).

AI-based systems are also important for innovation and are related to the areas of analytics and Big Data processing. One of the most advanced projects in this area is IBM Watson Analytics (see Chapter 6). For comprehensive coverage of AI, including definitions and its history, frontiers, and future, see Kaplan (2016).

Note: In January 2016, Mark Zuckerberg, the CEO of Facebook, announced publicly that his goal for 2016 was to build an AI-based assistant to help with his personal and business activities and decisions. Zuckerberg was teaching a machine to understand his voice and follow his basic commands as well as to recognize the faces of his friends and business partners. Personal assistants are used today by millions of people (see Chapter 12).

Example: Pitney Bowes Is Getting Smarter with AI

Pitney Bowes Inc. is a U.S.-based global business solutions provider in areas such as product shipments, location intelligence, customer engagement, and customer information management. The company powers billions of physical and digital transactions annually across the connected and borderless world of commerce.

Today, at Pitney Bowes, shipping prices are determined automatically based on the dimensions, weight, and packaging of each package. The fee calculations create data that are fed into AI algorithms. The more data processed, the more accurate are the calculations (a machine-learning characteristic). The company estimates a 25 percent improvement in calculations achieved from their algorithms. This gives Pitney Bowes an accurate base for pricing, better customer satisfaction, and improved competitive advantage.

Major Goals of AI

The overall goal of AI is to create intelligent machines that are capable of executing a variety of tasks currently done by people. Ideally, AI machines should be able to reason, think abstractly, plan, solve problems, and learn.

Some specific goals are to:

- Perceive and properly react to changes in the environment that influence specific business processes and operations.
- Introduce creativity in business processes and decision making.

Drivers of AI

The use of AI has been driven by the following forces:

- People's interest in smart machines and artificial brains
- The low cost of AI applications versus the high cost of manual labor (doing the same work)
- The desire of large tech companies to capture competitive advantage and market share of the AI market and their willingness to invest billions of dollars in AI
- The pressure on management to increase productivity and speed
- The availability of quality data contributing to the progress of AI
- The increasing functionalities and reduced cost of computers in general
- The development of new technologies, particularly cloud computing

Benefits of AI

The major benefits of AI are as follows:

- AI has the ability to complete certain tasks much faster than humans.
- The consistency of the completed AI work can be much better than that of humans. AI machines do not make mistakes.
- AI systems allow for continuous improvement projects.
- AI can be used for predictive analysis via its capability of pattern recognition.
- AI can manage delays and blockages in business processes.
- AI machines do not stop to rest or sleep.
- AI machines can work autonomously or be assistants to humans.
- The functionalities of AI machines are ever increasing.
- AI machines can learn and improve their performance.
- AI machines can work in environments that are hazardous to people.
- AI machines can facilitate innovations by human (i.e., support research and development [R&D]).
- No emotional barriers interfere with AI work.
- AI excels in fraud detection and in security facilitations.
- AI improves industrial operations.
- AI optimizes knowledge work.
- AI increases speed and enables scale.
- AI helps with the integration and consolidating of business operations.
- AI applications can reduce risk.
- AI can free employees to work on more complex and productive jobs.
- AI improves customer care.
- AI can solve difficult problems that previously were unsolved (Kharpal, 2017).
- AI increases collaboration and speeds up learning.

These benefits facilitate competitive advantage as reported by Agrawal (2018).

Note: Not all AI systems deliver all these benefits. Specific systems may deliver only some of them.

The capability of reducing costs and increasing productivity may result in large increases in profit (Violino, 2017). In addition to benefiting individual companies, AI can dramatically increase a country's economic growth, as it is doing in Singapore.

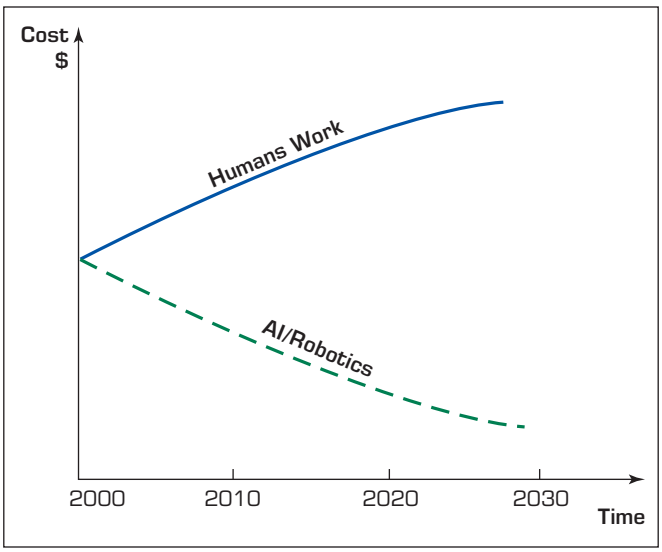


FIGURE 2.2 **Cost of Human Work versus the Cost of AI Work.**

EXAMPLES OF AI BENEFITS The following are typical benefits of AI in various areas of applications:

1. The International Swabs and Derivatives Association (ISDA) uses AI to eliminate tedious activities in contract procedures. For example, by using optical character recognition (OCR) integrated with AI, ISDA digitizes contracts and then defines, extracts, and archives the contracts.
2. AI is starting to revolutionize business recruitment by (1) conducting more efficient and fairer candidate screening, (2) making better matches of candidates to jobs, and (3) helping safeguard future talent pipelines for organizations. For details, see SMBWorld Asia Editors (2017) and Section 2.8.
3. AI is redefining management. According to Kolbjørnsrud et al. (2016), the following five practices result from the use of AI:
   - It can perform routine administrative tasks.
   - Managers can focus on the judgment portions of work.
   - Intelligent machines are treated as colleagues (i.e., managers trust the advice generated by AI). In addition, there is people–machine collaboration (see Chapter 11).
   - Managers concentrate on creative abilities that can be supported by AI machines.
   - Managers are developing social skills, which are needed for better collaboration, leadership, and coaching.
4. Accenture Inc. developed AI-powered solutions using natural language processing (NLP) and image recognition to help blind people in India improve the way that they can experience the world around them. This enables them to have a better life, and those who work can work better, faster, and do jobs that are more challenging.
5. Ford Motor Credit uses machine learning to spot overlooked borrowers. In addition, it uses machine learning to help its underwriters better understand loan applicants. The program helps the productivity of both underwriters and overlooked applicants. Finally, the system predicts potential borrowers' creditworthiness, thus minimizing losses for Ford.
6. Alastair Cole uses data collected from several sources with IBM Watson to predict what customers are expecting from the company. The generated data are used for supporting more efficient business decisions.
7. Companies are building businesses around AI. There are many examples of start-ups or existing companies that are attempting to create new businesses.

Two areas in which large benefits have already been reaped are customer experience and enjoyment. According to a global survey reported by CMO Innovation Editors (2017), 91 percent of top-performing companies deployed AI solutions to support customer experience.

Some Limitations of AI Machines

The following are the major limitations of AI machines:

- Lack human touch and feel
- Lack attention to non-task surroundings
- Can lead people to rely on AI machines too much (e.g., people may stop to think on their own)
- Can be programmed to create destruction (see discussion in Chapter 14)
- Can cause many people to lose their jobs (see Chapter 14)
- Can start to think by themselves, causing significant damage (see Chapter 14)

Some of the limitations are diminishing with time. However, risks exist. Therefore, it is necessary to properly manage the development of AI and try to minimize risks.

WHAT AI CAN AND CANNOT DO The limitations just identified constrain the capabilities of commercial AI. For example, it could cost too much to be commercially used. Ng (2016) provides an assessment of what AI was able to do by 2016. This is important for two reasons: (1) executives need to know what AI can do economically and how companies can use it to benefit their business and (2) executives need to know what AI cannot economically do.

AI is already transforming Web search, retailing and banking services, logistics, online commerce, entertainment, and more. Hundreds of millions of people use AI on their smartphones and in other ways. However, according to Ng (2016), applications in these areas are based on how simple input is converted to simple output as a response; for example, in automatic loan approval, the input is the profile of the applicant and the output will be an approval or rejection.

Applications in these areas are normally fully automated. Automated tasks are usually repetitive and done by people with short periods of training. AI machines depend on data that may be difficult to get (e.g., belong to someone else) or inaccurate. A second barrier is the need for AI experts, who are difficult to find and/or expensive to hire. For other barriers, see Chapter 14.

Three Flavors of AI Decisions

Staff (2017) divided the capabilities of AI systems into three levels: assisted, autonomous, and augmented.

ASSISTED INTELLIGENCE This is equivalent mostly to weak AI, which works only in narrow domains. It requires clearly defined inputs and outputs. Examples are some monitoring systems and low-level virtual personal assistants (Chapter 12). Such systems and assistants are used in our vehicles for giving us alerts. Similar systems can be used in many healthcare applications (e.g., monitoring, diagnosing).

AUTONOMOUS AI These systems are in the realm of the strong AI but in a very narrow domain. Eventually, a computer will take over many tasks, automating them completely. Machines act as experts and have absolute decision-making power. Pure robo-advisors (Chapter 12) are examples of such machines. Autonomous vehicles and robots that can fix themselves are also good examples.

AUGMENTED INTELLIGENCE Most of the existing AI applications, which are between assisted and autonomous, are referred to as **augmented intelligence** (or intelligence augmentation). Their technology can augment computer tasks to extend human cognitive abilities (see Chapter 6 on cognitive computing), resulting in high performance, as described in Technology Insight 2.1.

Artificial Brain

The **artificial brain** is a people-made machine that is desired to be as intelligent, creative, and self-aware as humans. To date, no one has been able to create such a machine; see **artificialbrains.com**. A leader in this area is IBM. IBM and the U.S. Air Force have built a system equivalent to 64 million artificial neurons that aims to reach 10 billion neurons

TECHNOLOGY INSIGHT 2.1 Augmented Intelligence

The idea of combining the performance of people and machines is not new. In this section, we discuss combining (augmenting) human abilities with powerful machine intelligence—not replacing people, which autonomous AI does, but extending human cognitive abilities. The result is the ability of humans to solve more complex problems, as in the opening vignette to Chapter 1. Computers have provided data to help people solve problems for which no solution had been available. Padmanabhan (2018) specifies the following differences between traditional and augmented AI:

1. Augmented machines extend human thinking capabilities rather than replace human decision making. These machines facilitate creativity.
2. Augmentation excels in solving complex human and industry problems in specific domains in contrast with strong, general AI machines, which are still in development.
3. In contrast with a "black box" model of some AI and analytics, the augmented intelligence provides insights and recommendations, including explanations.
4. In addition, augmented technology can offer new solutions by combining existing and discovered information in contrast to assisted AI that identifies problems or symptoms and suggests predetermined known solutions.

Padmanabhan (2018) and many others believe that at the moment, augmented AI is the best option to deal with practical problems and transform organizations to be "smarter."

In contrast with autonomous AI, which describes machines with a wide range of cognitive abilities (e.g., driverless cars), augmented intelligence has only a few cognitive abilities.

Examples of Augmented Intelligence

Staff (2017) provides the following areas for which AI is useful:

- **Cybercrime fighting.** For example, AI can identify forthcoming attacks and suggest solutions.
- **E-commerce decisions.** AI marketing tools can make testing results 100 times faster, and adapt the layout and response functions of a Web site to users. Machines also make recommendations, and marketers can accept or reject them.
- **High-frequency stock market trading.** This process can be done either completely autonomously or in some cases with human control and calibration.

Questions for Discussion

1. What is the basic premise of augmented intelligence?
2. List the major differences between augmented intelligence and assisted AI applications.
3. What are some benefits of augmented intelligence?
4. How does the technology relate to cognitive computing?

by 2020. Note that a human brain contains about 100 billion neurons. The system tries to imitate a biological brain and be energy efficient. IBM's project is called TrueNorth or BlueBrain, and it learns from humans' brains. Many believe that it will be a long and slow process for AI machines to be as creative as people (e.g., Dormehl, 2017).

SECTION 2.2 REVIEW QUESTIONS

1. Define AI.
2. What are the major aims and goals of AI?
3. List some characteristics of AI.
4. List some AI drivers.
5. List some benefits of AI applications.
6. List some AI limitations.
7. Describe the artificial brain.
8. List the three flavors of AI and describe augmentation.

2.3 HUMAN AND COMPUTER INTELLIGENCE

AI usage is growing rapidly due to its increased capabilities. To understand AI, we need to first explore the meaning of intelligence.

What Is Intelligence?

Intelligence can be considered to be an umbrella term and is usually measured by an IQ test. However, some claim that there are several types of intelligence. For example, Dr. Howard Gardner of Harvard University proposed the following types of intelligence:

- Linguistic and verbal
- Logical
- Spatial
- Body/movement
- Musical
- Interpersonal
- Intrapersonal
- Naturalist

Thus, intelligence is not a simple concept.

CONTENT OF INTELLIGENCE Intelligence is composed of reasoning, learning, logic, problem-solving ability, perception, and linguistic ability.

Obviously, the concept of intelligence is not simple.

CAPABILITIES OF INTELLIGENCE To understand what artificial intelligence is, it is useful to first examine those abilities that are considered signs of human intelligence:

- Learning or understanding from experience
- Making sense out of ambiguous, incomplete, or even contradictory messages and information
- Responding quickly and successfully to a new situation (i.e., using the most correct responses)
- Understanding and inferring in a rational way, solving problems, and directing conduct effectively

- Applying knowledge to manipulate environments and situations
- Recognizing and judging the relative importance of different elements in a situation

AI attempts to provide some, hopefully all, of these capabilities, but in general, it is still not capable of matching human intelligence.

How Intelligent Is AI?

AI machines have demonstrated superiority over humans in playing complex games such as chess (beating the world champion), *Jeopardy!* (beating the best players), and Go (a complex Chinese game) whose top players were beaten by a computer using the well-known program, Google's DeepMind (see Hughes, 2016). Despite these remarkable demonstrations (whose cost is extremely high), many AI applications still show significantly less intelligence than humans.

COMPARING HUMAN INTELLIGENCE WITH AI Several attempts have been made to compare human intelligence with AI. There is difficulty in doing so because it is a multidimensional situation. A comparison is presented in Table 2.1.

TABLE 2.1 Artificial Intelligence versus Human Intelligence

Area	AI	Human
Execution	Very fast	Can be slow
Emotions	Not yet	Can be positive or negative
Computation speed	Very fast	Slow, may have trouble
Imagination	Only what is programmed for	Can expand existing knowledge
Answers to questions	What is in the program	Can be innovative
Flexibility	Rigid	Large, flexible
Foundation	A binary code	Five senses
Consistency	High	Variable, can be poor
Process	As modeled	Cognitive
Form	Numbers	Signals
Memory	Built in, or accessed in the cloud	Use of content and scheme memory
Brain	Independent	Connected to a body
Creativity	Uninspired	Truly creative
Durability	Permanent, but can get obsolete if not updated	Perishable, but can be updated
Duplication, documentation, and dissemination	Easy	Difficult
Cost	Usually low and declining	Maybe high and increasing
Consistency	Stable	Erratic at times
Reasoning process	Clear, visible	Difficult to trace at times
Perception	By rules and data	By patterns
Figure missing data	Usually cannot	Frequently can

For additional comparisons and who had the advantage in which area, see **www.dennisgorelik.com/ai/ComputerintelligenceVsHumanIntelligence.htm**.

Measuring AI

The Turing Test is a well-known attempt to measure the intelligence level of AI machines.

TURING TEST: THE CLASSICAL MEASURE OF MACHINE INTELLIGENCE Alan Turing designed a test known as the **Turing Test** to determine whether a computer exhibits intelligent behavior. According to this test, a computer can be considered smart only when a human interviewer asking the same questions to both an unseen human and an unseen computer cannot determine which is which (see Figure 2.3). Note that this test is limited to a question-and-answer (Q&A) mode.

To pass the Turing Test, a computer needs to be able to understand a human language (NLP), to possess human intelligence (e.g., have a knowledge base), to reason using its stored knowledge, and to be able to learn from its experiences (machine learning).

Note: The $100,000 Leobner prize is waiting for the person or persons who develop software that is truly intelligent (i.e., passing the Turing Test).

OTHER TESTS Over the years, there have been several other proposals of how to measure machine intelligence. For example, improvements in the Turing Test appear in several variants. Major U.S. universities (e.g., University of Illinois, Massachusetts Institute of Technology [MIT], Stanford University) are engaged in studying the IQ of AI. In addition, there are several other measuring tests. Let's examine one test in Application Case 2.1.

In conclusion, it is difficult to measure the level of intelligence of humans as well as that of machines. Doing so depends on the circumstances and the metrics used.

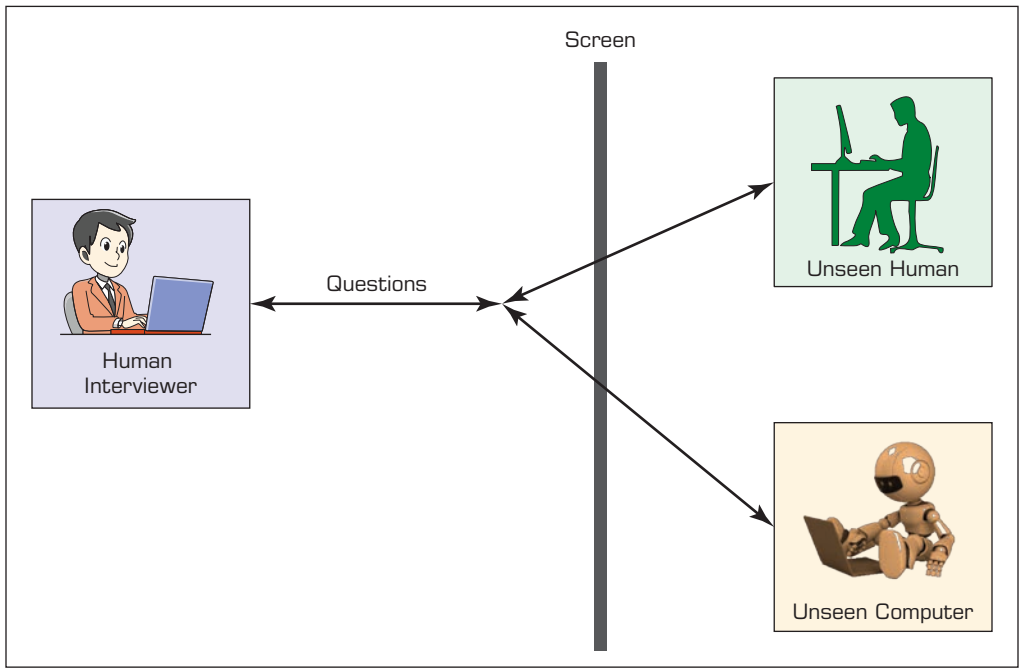


FIGURE 2.3 A Pictorial Representation of the Turing Test

Application Case 2.1 How Smart Can a Vacuum Cleaner Be?

If you do not know it, vacuum cleaners can be smart. Some of you may use the Roomba from iRobot. This vacuum cleaner can be left alone to clean floors, and it exhibits some intelligence.

However, in smart homes (Chapter 13), we expect to see even smarter vacuum cleaners. One is Roboking Turbo Plus from LG in Korea. Researchers at South Korea's Seoul National University Robotics and Intelligent System Lab studied the Roboking and verified that its deep-learning algorithm makes it as intelligent as a six- or seven-year-old child. If we have self-driving cars, why can't we have a self-driving vacuum cleaner, which is much simpler than a car. The cleaner needs only to move around an entire room. To do so, the machine needs to "see" its location in a room and identify obstacles in front of it. Then the cleaner's knowledge base needs to find what is the best thing to do (given worked in the past). This is basically what many AI machines' sensors, knowledge bases, and rules do. In addition, the AI machine needs to learn from its past experience (e.g., what it should not do because it did not work in the past).

Roboking is equipped with LG's Deep Thin QTM AI program, which enables the vacuum cleaner to figure out the nature of an encountered obstacle. The program tells it to go around furniture, wait for a dog to move, or stop. So, how intelligent is the machine? To answer this question, the Korean researchers developed 100 metrics and tested vacuum cleaners that were boasted as autonomous. The performance of the tested cleaners was divided into three levels based on their performance regarding the 100 metrics. The levels were as intelligent as a dolphin, as intelligent as an ape, and as intelligent as a six-to-seven-year-old child. The study confirmed that Roboking performed tasks at the upper level of machine intelligence.

Sources: Compiled from Fuller (2017) and **webwire.com/ViewPressRel.asp?aId=211017** news dated July 18, 2017.

Questions for Case 2.1

1. How did the Korean researchers determine the performance of the vacuum cleaners?
2. If you own (or have seen) the Roomba, how intelligent do you think it is?
3. What capability can be generated by the deep learning feature? (You need to do some research.)
4. Find recent information about LG's Roboking. Specifically, what are the newest improvements to the product?

Regardless of the determination of how intelligent a machine is, AI exhibits a large number of benefits as described earlier.

It is important to note that the capabilities of AI are increasing with time. For example, an experiment at Stanford University (Pham, 2018) found that AI programs at Microsoft and Alibaba Co. have scored higher than hundreds of individual people at reading comprehension tests. (Of course, these are very expensive AI programs.) For a discussion of AI versus human intelligence, see Carney (2018).

SECTION 2.3 REVIEW QUESTIONS

1. What is intelligence?
2. What are the major capabilities of human intelligence? Which are superior to that of AI machines?
3. How intelligent is AI?
4. How can we measure AI's intelligence?
5. What is the Turing Test and what are its limitations?
6. How can one measure the intelligence level of a vacuum cleaner?

2.4 MAJOR AI TECHNOLOGIES AND SOME DERIVATIVES

The AI field is very broad; we can find AI technologies and applications in hundreds of disciplines ranging from medicine to sports. Press (2017) lists 10 top AI technologies similar to what is covered in this book. Press also provides the status of the life cycle (ecosystem phase) of the technologies. In this section, we present some major AI technologies and their derivatives as related to business. The selected list is illustrated in Figure 2.4.

Intelligent Agents

An **intelligent agent (IA)** is an autonomous, relatively small computer software program that observes and acts upon changes in its environment by running specific tasks autonomously. An IA directs an agent's activities to achieve specific goals related to the changes in the surrounding environment. Intelligent agents may have the ability to learn by using and expanding the knowledge embedded in them. Intelligent agents are effective tools for overcoming the most critical burden of the Internet information overload and making computers more viable decision support tools. Interest in using intelligent agents for business and e-commerce started in the academic world in the mid-1990s. However, only since 2014, when the capabilities of IA increased remarkably, have we started to see powerful applications in many areas of business, economics, government, and services.

Initially, intelligent agents were used mainly to support routine activities such as searching for products, getting recommendations, determining products' pricing, planning marketing, improving computer security, managing auctions, facilitating payments, and improving inventory management. However, these applications were very simple, using a low level of intelligence. Their major benefits were increasing speed, reducing costs, reducing errors, and improving customer service. Today's applications, as we will see throughout this chapter, are much more sophisticated.

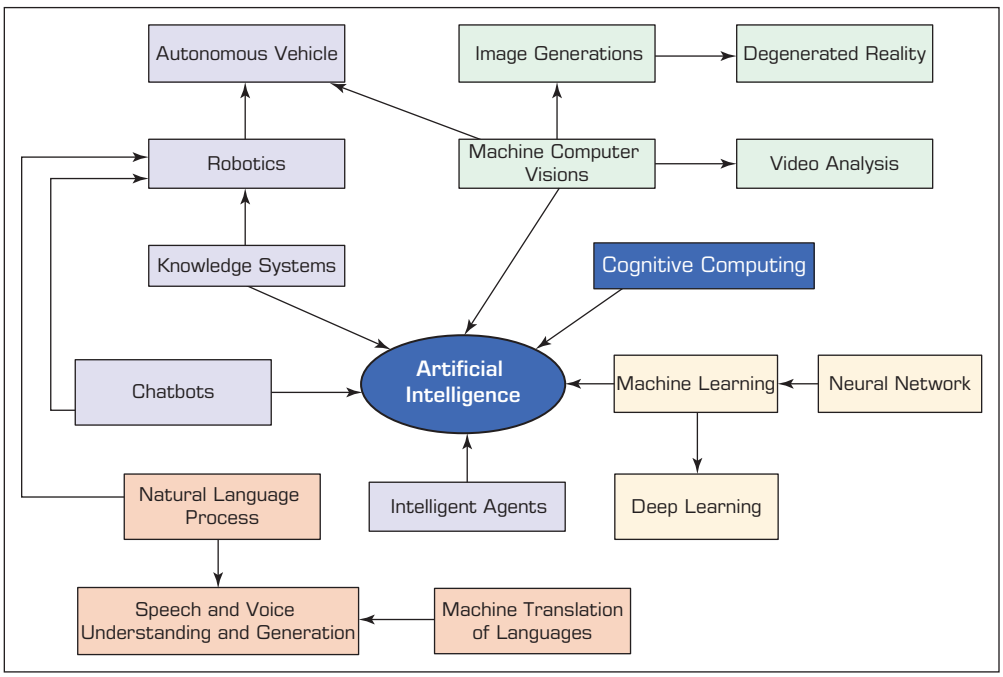


FIGURE 2.4 The Major AI Technologies

Example 1: Virus Detection Program

A simple example of an intelligent software agent is a virus detection program. It resides in a computer, scans all incoming data, and removes found viruses automatically while learning to detect new virus types and detection methods.

Example 2

Allstate Business Insurance is using an intelligent agent to reduce call center traffic and provide help to human insurance agents during the rate-quoting process with business customers. In these cases, rate quotes can be fairly complicated. Using this system, agents can quickly answer questions posted by corporate customers, even if the agents are not fully familiar with the related issue.

Intelligent agents are also utilized in e-mail servers, news filtering and distribution, appointment handling, and automated information gathering.

Machine Learning

At this time, AI systems do not have the same learning capabilities that humans have; rather, they have simplistic (but improving) **machine learning** (modeled after human learning methods. The machine-learning scientists try to teach computers to *identify patterns* and make connections by showing the machines a large volume of examples and related data. Machine learning also allows computer systems to monitor and sense their environmental activities so the machines can adjust their behavior to deal with changes in the environment. The technology can also be used to predict performance, to reconfigure programs based on changing conditions, and much more. Technically speaking, machine learning is a scientific discipline concerned with the design and development of algorithms that allow computers to learn based on data coming from sensors, databases, and other sources. This learning is then used for making predictions, recognizing patterns, and supporting decision makers. For an overview, see Alpaydin (2016) and Theobald (2017).

Machine-learning algorithms (see Chapter 5 for description and discussion) are used today by many companies. For an executive guide to machine learning, see Pyle and San Jose (2015).

The process of machine learning involves computer programs that learn as they face new situations. Such programs collect data and analyze them and then "train" themselves to arrive at conclusions. For example, by showing examples of situations to a machine-learning program, the program can find elements not easily visible without it. A well-known example is that of computers detecting credit card fraud.

Application Case 2.2 illustrates how machine learning can improve companies' business processes.

According to Taylor (2016), the "increased computing power, coupled with other improvements including better algorithms and deep neural networks for image processing, and ultra-fast in-memory databases like SAP HANA, are the reasons why machine learning is one of the hottest areas of development in enterprise software today." Machine-learning applications are also expanding due to the availability of Big Data sources, especially those provided by the IoT (Chapter 13). Machine learning is basically learning from data.

There are several methods of machine learning. They range from neural networks to case-based reasoning. The major ones are presented in Chapter 5.

DEEP LEARNING One subset, or refinement, of machine learning is called **deep learning**. This technology, which is discussed in Chapter 6, tries to mimic how the human brain

Application Case 2.2 How Machine Learning Is Improving Work in Business

The following examples of using machine learning are provided by Wellers, et al. (2017), who stated that "today's leading organizations are using machine learning-based tools to automate decision processes. . . ."

1. *Improving customer loyalty and retention.* Companies mine customers' activities, transactions, and social interactions and sentiments to predict customer loyalty and retention. Companies can use machine learning, for example, to predict people's desire to change jobs and then employers can make attractive offers to keep the existing employees or to lure potential employees who work elsewhere to move to new employers.
2. *Hiring the right people.* Given an average of 250 applicants for a good job in certain companies, an AI-based program can analyze applicants' resumes and find qualified candidates who did not apply but placed their resume online.
3. *Automating finance.* Incomplete financial transactions that lack some data (e.g., order numbers) require special attention. Machine-learning systems can learn how to detect and correct such situations, very quickly and at minimal cost. The AI program can take the necessary corrective action automatically.
4. *Detecting fraud.* Machine-learning algorithms use pattern recognition to detect fraud in real time. The program is looking for anomalies, and then it makes inferences regarding the type of detected activities to look for fraud. Financial institutions are the major users of this program.
5. *Providing predictive maintenance.* Machine learning can find anomalies in the operation of equipment before it fails. Thus, corrective actions are done immediately at a fraction of a cost to repair equipment after it fails. In addition, optimal preventive maintenance can be done (see Opening Vignette Chapter 1).
6. *Providing retail shelf analysis.* Machine learning combined with machine vision can analyze displays in physical stores to find whether items are in proper locations on the shelves, whether the shelves are properly stocked, and whether the product labels (including prices) are properly shown.
7. *Making other predictions.* Machine learning has been used for making many types of predictions ranging in areas from medicine to investments. An example is Google Flights, which predicts delays that have not been flagged yet by the airlines.

Source: Compiled from Wellers, et al. (2017) and Theobald (2017).

Questions for Case 2.2

1. Discuss the benefits of combining machine learning with other AI technologies.
2. How can machine learning improve marketing?
3. Discuss the opportunities of improving human resource management.
4. Discuss the benefits for customer service.

works. Deep learning uses artificial neural technology and plays a major role in dealing with complex applications that regular machine learning and other AI technologies cannot handle. Deep learning (DL) delivers systems that not only think but also keep learning, enabling self-direction based on fresh data that flow in. DL can tackle previously unsolvable problems using its powerful learning algorithms.

For example, DL is a key technology in autonomous vehicles by helping to interpret road signs and road obstacles. DL is also playing critical roles in smartphones, robotics, tablets, smart homes, and smart cities (Chapter 13). For a discussion of these and other applications, see Mittal (2017). DL is mostly useful in real-time interactive applications in the areas of machine vision, scene recognition, robotics, and speech and voice processing. The key is continuous learning. As long as new data arrive, learning occurs.

Example

Cargill Corp. offers conventional analytics, and DL-based analytics help farmers to do more profitable work. For example, farmers can produce better shrimp at lower cost. DL is used extensively in stock market analysis and predictions. For details, see Smith (2017) and Chapter 6.

Machine and Computer Vision

The definitions of **machine vision** vary because several different computer vision systems include different hardware and software as well as other components. Generally speaking, the classical definition is that the term *machine vision* includes "the technology and methods used to provide imaging-based automated inspection and analysis for applications such as robot guidance, process control, autonomous vehicles, and inspection." Machine vision is an important tool for the optimization of production and robotic processes. A major part of machine vision is the industrial camera, which captures, stores, and archives visual information. This information is then presented to users or computer programs for analysis and eventually for automatic decision making or for support of human decision making. Machine vision can be confused with computer vision because sometimes the two are used as synonyms, but some users and researchers treat them as different entities. Machine vision is treated more as an engineering subfield, while computer vision belongs to the computer science area.

COMPUTER VISION **Computer vision**, according to Wikipedia, "is an interdisciplinary field that deals with how computers can be made for gaining high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do." Computer vision acquires or processes, analyzes, and interprets digital images and produces meaningful information for making decisions. Image data can take several formats, such as photos or videos, and they can come from multidimensional sources (e.g., medical scanners). Scene and item recognitions are important elements in computer vision. The computer vision field plays a vital role in the domains of safety, security, health, and entertainment. Computer vision is considered a technology of AI, which enables robots and autonomous vehicles to see (refer to the description in Chapter 6). Both computer vision and machine vision automate many human tasks (e.g., inspection). These tasks can deal with one image or a sequence of images. The major benefit of both technologies is lowering the costs of performing tasks, especially those that are repetitive and make the human eyes tired. The two technologies are also combined with *image processing* that facilitates complex applications, such as in visual quality control. Another view shows them as being interrelated based on image processing and sharing a variety of contributing fields.

An applied area of machine vision is **scene recognition**, which is performed by computer vision. Scene recognition enables recognition and interpretation of objects, scenery, and photos.

Example of Application

Significant illegal logging exists in many countries. To comply with the laws in the United States, Europe, and other countries, it is necessary to examine wood in the field. This requires expertise. According to the U.S. Department of Agriculture, "the urgent need for such field expertise, training and deploying humans to identify processed wood in the field [i.e., at ports, border crossings, weigh-stations, airports, and other points of entry for commerce] is prohibitively expensive and difficult logistically. The machine vision wood

identification project (MV) has developed a prototype machine vision system for wood identification." Similarly, AI computer vision combined with deep learning is used to identify illegal poachers of animals (see USC, 2018).

Another example of this application is *facial recognition* in several security applications, such as those used by the Chinese police that employ smart glasses to identify (via facial recognition) potential suspects. In 2018, the Chinese police identified a suspect who attended a pop concert. There were 60,000 people in the crowd. The person was recognized at the entrance gate where a camera took his picture; see the video at **youtube.com/watch?v=Fq1SEqNT-7c**. In 2018, US Citizenship and Immigration Services identified people that used false passports in the same manner.

VIDEO ANALYTICS Applying computer vision techniques to videos enables the recognition of patterns (e.g., for detecting fraud) and identifying events. This is a derivative application of computer vision. Another example is one in which, by letting computers view TV shows, it is possible to train the computers to make predictions regarding human interactions and the success of advertising.

Robotic Systems

Sensory systems, such as those for scene recognition and signal processing, when combined with other AI technologies, define a broad category of integrated, possibly complex, systems, generally called *robotics* (Chapter 10). There are several definitions of robots, and they are changing over time. A classical definition is this: "A **robot** is an electromechanical device that is guided by a computer program to perform manual and/or mental tasks." The Robotics Institute of America formally defines a robot as "a programmable multifunctional manipulator designed to move materials, parts, tools, or specialized devices through variable programmed motions for the performance of a variety of tasks." This definition ignores the many mental tasks done by today's robots.

An "intelligent" robot has some kind of sensory apparatus, such as a camera, that collects information about the robot's surroundings and its operations. The collected data are interpreted by the robot's "brain," allowing it to respond to the changes in the environment.

Robots can be fully autonomous (programmed to do tasks completely on their own, even repair themselves), or can be remotely controlled by a human. Some robots known as *androids* resemble humans, but most industrial robots are not this type. Autonomous robots are equipped with AI intelligent agents. The more advanced smart robots are not only autonomous but also can learn from their environment, building their capabilities. Some robots today can learn complex tasks by watching what humans do. This leads to better human–robot collaboration. The Interactive Group at MIT is experimenting with this capability by teaching robots to make complex decisions. For details, see Shah (2016). For an overview of the robot revolution, see Waxer (2016).

Example: Walmart Is Using Robots to Properly Stock Shelves

The efficiency of Walmart stores depends on appropriately stocking their shelves. Using manual labor for checking what is going on is expensive and may be inaccurate. As of late 2017, robots were supporting the company's stocking decisions.

At Walmart, the 2-foot-tall robots use a camera/sensor to scan the shelves to look for misplaced, missing, or mispriced items. The collected information and the interpretation of problems are done by these self-moving robots. The results are transmitted to humans who take corrective actions. The robots carry out their tasks faster and frequently more accurately than humans. The company experimented with this in 50 stores in 2018.

Preliminary results are significantly positive and are also expected to increase customer satisfaction. The robots will not cause employees to lose their jobs.

Robots are used extensively in e-commerce warehouses (e.g., tens of thousands are used by **Amazon.com**). They also are used in make-to-order manufacturing as well as in mass production (e.g., cars), lately of self-driven vehicles. A new generation of robots is designed to work as advisors, as described in Chapter 12. These robots are already advising on topics such as investments, travel, healthcare, and legal issues. Robots can serve as front desk receptionists and even can be used as teachers and trainers.

Robots can help with online shopping by collecting shopping information, matching buyers and products, and conducting price and capability comparisons. These are known as **shopbots** (e.g., see **igi-global.com/dictionary/shopbot/26826**). Robots can carry goods for shoppers in open air markets. Walmart is experimenting now with robotic shopping carts (Knight, 2016). For a video (4:41 min.), see **businessinsider.com/personal-robots-for-shopping-and-e-commerce-2016-9?IR=T**. The Japanese company SoftBank opened a cellphone store in Tokyo entirely staffed by robots, each named Pepper. Each robot is mobile (on wheels) and can approach customers. Initially, communication with customers was done by entering information into a tablet attached to each Pepper. A major issue with robots is their trend to take human jobs. For a discussion of this topic, see Section 14.6.

Natural Language Processing

Natural language processing (NLP) is a technology that gives users the ability to communicate with a computer in their native language. The communication can be in written text and/or in voice (speech). This technology allows for a conversational type of interface in contrast with using a programming language that consists of computer jargon, syntax, and commands. NLP includes two subfields:

- *Natural language understanding* that investigates methods of enabling computers to comprehend instructions or queries provided in ordinary English or other human languages.
- *Natural language generation* that strives to have computers produce ordinary spoken language so that people can understand the computers more easily. For details and the history of NLP, see **en.wikipedia.org/wiki/Natural_language_processing** and Chapter 6.

NLP is related to voice-generated data as well as text and other communication forms.

SPEECH (VOICE) UNDERSTANDING **Speech (voice) understanding** is the recognition and understanding of spoken languages by a computer. Applications of this technology have become more popular. For instance, many companies have adopted this technology in their automated call centers. For an interesting application, see **cs.cmu.edu/~./listen**.

Related to NLP is machine translation of languages, which is done by both written text (e.g., Web content) and voice conversation.

MACHINE TRANSLATION OF LANGUAGES Machine translation uses computer programs to translate words and sentences from one language to another. For example, Babel Fish Translation, available at **babelfish.com**, offers more than 25 different combinations of language translations. Similarly, Google's Translate (**translate.google.com**) can translate dozens of different languages. Finally, users can post their status on Facebook in several languages.

Example: Sogou's Travel Translator

This Chinese company introduced, in 2018, an AI-powered portable travel device. Chinese people are now traveling to other countries in increasing numbers (200 million expected in 2020 versus 122 million in 2016). The objective of the device is to enable Chinese tourists to plan trips (so they can read Web sites like Trip Advisor, available in English). The AI-powered portable travel device enables tourists to read menus, street signs, and communicate with native speakers. The device, which is using NLP and image recognition, is connected to Sogou search (a search engine). In contrast with the regular Chinese-English dictionaries, this device is structured specifically for travelers and their needs.

Knowledge and Expert Systems and Recommenders

These systems, which are presented in Chapter 12, are computer programs that store knowledge, which their applications use to generate expert advice and/or perform problem solving. Knowledge-based expert systems also help people to verify information and make certain types of automated routine decisions.

Recommendation systems (Chapter 12) are knowledge-based systems that make shopping and other recommendations to people. Another knowledge system is chatbots (see Chapter 12).

KNOWLEDGE SOURCES AND ACQUISITION FOR INTELLIGENT SYSTEMS For many intelligent systems to work, it is necessary for them to have knowledge. The process of acquiring this knowledge is referred to as **knowledge acquisition**. This activity can be complex because it is necessary to make sure what knowledge is needed. It must fit the desired system. In addition, the sources of the knowledge need to be identified to ensure the feasibility of acquiring the knowledge. The specific methods of acquiring the knowledge need to be identified and if expert(s) are the source of knowledge, their cooperation must be ensured. In addition, the method of knowledge representation and reasoning from the collected knowledge must be taken into account, and knowledge must be validated and be consistent.

Given this information, it is easy to see that the process of knowledge acquisition (see Figure 2.5) can be very complex. It includes extracting and structuring knowledge. It has several methods (e.g., observing, interviewing, scenario building, and discussing), so specially trained knowledge engineers may be needed for knowledge acquisition and system building. In many cases, teams of experts with different skills are created for knowledge acquisition. Knowledge can be generated from data, and then experts may be used to verify it. The acquired knowledge needs to be organized in an activity referred to as *knowledge representation*.

KNOWLEDGE REPRESENTATION Acquired knowledge needs to be organized and stored. There are several methods of doing this, depending on what the knowledge will be used for, how the reasoning from this knowledge will be done, how users will interact with the knowledge, and more. A simple way to represent knowledge is in the form of questions and matching answers (Q&A).

REASONING FROM KNOWLEDGE Perhaps the most important component in an intelligent system is its reasoning feature. This feature processes users' requests and provides answers (e.g., solutions, recommendations) to the user. The major difference among the various types of the intelligent technologies is the type of reasoning they use.

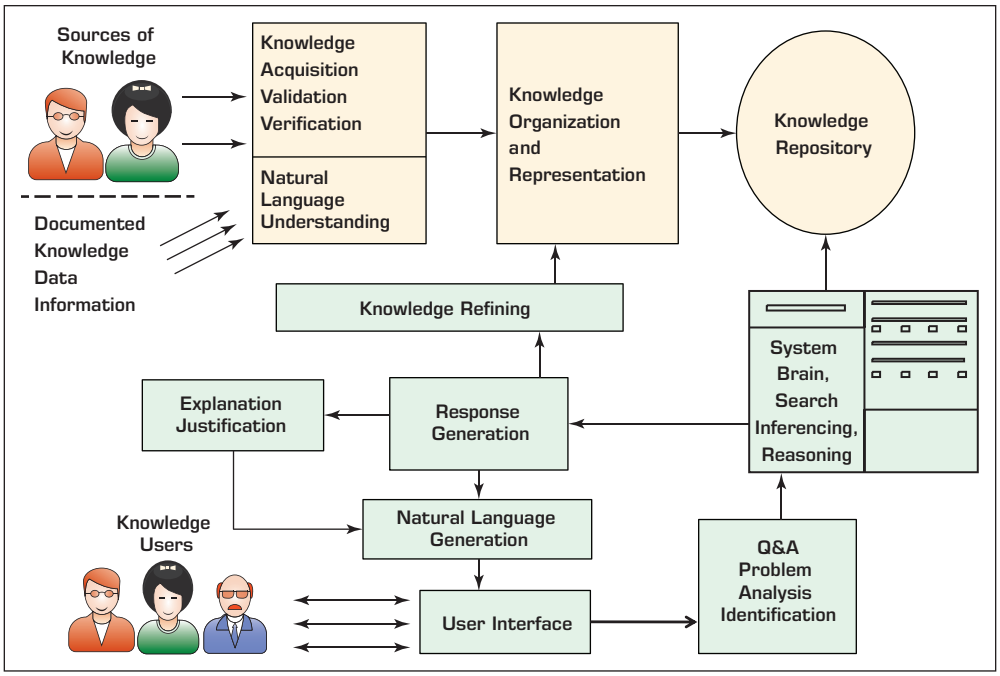


FIGURE 2.5 Automated Decision-Making Process

Chatbots

Robots come in several shapes and types. One type that has become popular in recent years is the chatbot. A chatbot, which will be presented in Chapter 12, is a conversional robot that is used for chatting with people. (A "bot" is short for "robot.") Depending on the purpose of the chat, which can be done in writing or by voice, bots can be in the form of intelligent agents that retrieve information or personal assistants that provide advice. In either case, chatbots are usually equipped with NLP that enables conversations in natural human languages rather than in a programmed computer language. Note that Google has rolled out six different voices to its Google's Assistant.

Emerging AI Technologies

Several new AI technologies are emerging. Here are a few examples:

- *Effective computing.* These technologies detect the emotional conditions of people and suggest how to deal with discovered problems
- *Biometric analysis. These technologies* can verify an identity based on unique biological traits that are compared to stored ones (e.g., facial recognition).

COGNITIVE COMPUTING **Cognitive computing** is the application of knowledge derived from cognitive science (the study of the human brain) and computer science theories in order to simulate the human thought processes (an AI objective) so that computers can exhibit and/or support decision-making and problem-solving capabilities (see Chapter 6). To do so, computers must be able to use *self-learning algorithms*, pattern recognition, NLP, machine vision, and other AI technologies. IBM is a major proponent of the concept by developing technologies (e.g., Watson) that support people in making complex decisions. Cognitive computing systems learn to reason with purpose, and interact with people naturally. For details, see Chapter 6 and Marr (2016).

AUGMENTED REALITY **Augmented reality** (AR) refers to the integration of digital information with the user environment in real time (mostly vision and sound). The technology provides people real-world interactive experience with the environment. Therefore, information may change the way people work, learn, play, buy, and connect. Sophisticated AI programs may include machine vision, scene recognition, and gesture recognition. AR is available on iPhones as ARKit. (Also see Metz, 2017.)

These AR systems use data captured by sensors (e.g., vision, sound, temperature) to augment and supplement real-world environments. For example, if you take a photo of a house with your cellphone, you can immediately get the publicly available information about its configuration, ownership, and tax liabilities on your cellphone.

SECTION 2.4 REVIEW QUESTIONS

1. Define *intelligent agents* and list some of their capabilities.
2. Prepare a list of applications of intelligent agents.
3. What is machine learning? How can it be used in business?
4. Define *deep learning*.
5. Define *robotics* and explain its importance for manufacturing and transportation.
6. What is NLP? What are its two major formats?
7. Describe machine translation of languages. Why it is important in business?
8. What are knowledge systems?
9. What is cognitive computing?
10. What is augmented reality?

2.5 AI SUPPORT FOR DECISION MAKING

Almost since the inception of AI, researchers have recognized the opportunity of using it for supporting the decision-making process and for completely automating decision making. Jeff Bezos, the CEO of **Amazon.com**, said in May 2017 that AI is in a golden age, and it is solving problems that were once in the realm of science fiction (Kharpal, 2017). Bezos also said that **Amazon.com** is using AI in literally hundreds of applications, and AI is really of amazing assistance. **Amazon.com** has been using AI, for example, for product recommendations for over 20 years. The company also uses AI for product pricing, and as Bezos said, to solve many difficult problems. And indeed, since its inception, AI has been related to problem solving and decision making. AI technologies allow people to make better decisions. The fact is that AI can:

- Solve complex problems that people have not been able to solve. (Note that solving problems frequently involves making decisions.)
- Make much faster decisions. For example, Amazon makes millions of pricing and recommendation decisions, each in a split second.
- Find relevant information, even in large data sources, very fast.
- Make complex calculations rapidly.
- Conduct complex comparisons and evaluations in real time.

In a nutshell, AI can drive some types of decisions many times faster and more consistently than humans can. For details, watch the video at **youtube.com/watch?v=Dr9jeRy9whQ/**. The nature of decision making, especially nonroutine ones, as noted in Chapter 1, is complex. We discussed in Chapter 1 the fact that there are several types of decisions and several managerial levels of making them, and we looked at the typical process of making decisions. Making decisions, many of which are used for problem solving, requires intelligence and expertise. AI's aim is to provide both. As a

result, it is clear that using AI to facilitate decision making involves many opportunities, benefits, and variations. For example, AI can successfully support certain types of decision making and fully automate others.

In this section, we discuss some general issues of AI decision support. The section also distinguishes between *support of decision making* and *fully automating decision making*.

Some Issues and Factors in Using AI in Decision Making

Several issues determine the justification of using AI and its chance of success. These include:

- The nature of the decision. For example, routine decisions are more likely to be fully automated, especially if they are simple.
- The method of support, what technology(ies) is (are) used. Initially, automated decision supports were rule-based. Practically, expert systems were created to generate solutions to specific decision situations in well-defined domains. Another popular technology mentioned earlier was "recommender," which appeared with e-commerce in the 1990s. Today, there is an increased use of machine learning and deep learning. A related technology is that of pattern recognition. Today, attention is also given to biometric types of recognition.

For example, research continues to develop an AI machine that will interview people at airports, asking one or two questions, and then determining whether they are telling the truth. Similar algorithms can be used to vet refugees and other types of immigrants.

- *Cost-benefit and risk analyses.* These are necessary for making large-scale decisions, but computing these values may not be simple with AI models due to difficulties in measuring costs, risks, and benefits. For example, as we cited earlier, researchers used 100 metrics to measure the intelligence level of vacuum cleaners.
- *Using business rules.* Many AI systems are based on business or other types of rules. The quality of automated decisions depends on the quality of these rules. Advanced AI systems can learn and improve business rules.
- *AI algorithms.* There is an explosion in the number of AI algorithms that are the basis for automated decisions and decision support. The quality of the decisions depends on the input of the algorithms, which may be affected by changes in the business environment.
- *Speed.* Decision automation is also dependent on the speed within which decisions need to be made. Some decisions cannot be automated because it takes too much time to get all the relevant input data. On the other hand, manual decisions may be too slow for certain circumstances.

AI Support of the Decision-Making Process

Much AI support can be applied today to the various steps of the decision-making process. Fully automated decisions are common in routine situations and will be discussed in the next section. Here we follow the steps in the decision-making process described in Chapter 1.

PROBLEM IDENTIFICATION AI systems are used extensively in problem identification typically in diagnosing equipment malfunction and medical problems, finding security breaches, estimating financial health, and so on. Several technologies are used. For example, sensor-collected data are used by AI algorithms. Performance levels of machines are compared to standards, and trend analysis can point to opportunities or troubles.

GENERATING OR FINDING ALTERNATIVE SOLUTIONS Several AI technologies offer alternative solutions by matching problem characteristics with best practices or proven solutions stored in databases. Both expert systems and chatbots employ this approach. They can generate recommended solutions or provide several options from which to choose. AI tools such as case-based reasoning and neural computing are used for this purpose.

SELECTING A SOLUTION AI models are used to evaluate proposed solutions, for example, by predicting their future impact (predictive analysis), assessing their chance of success, or predicting a company's reply to action taken by a competitor.

IMPLEMENTING THE SOLUTIONS AI can be used to support the implementation of complex solutions. For example, it can be used to demonstrate the superiority of proposals and to assess resistance to changes.

Applying AI to one or more of the decision-making processes and steps enables companies to solve complex real-world problems, as shown in Application Case 2.3.

Automated Decision Making

As the power of AI technologies increases, so does its ability to fully automate more and more complex decision-making situations.

Application Case 2.3

How Companies Solve Real-World Problems Using Google's Machine-Learning Tools

The following examples were extracted from Forrest (2017):

Google's Cloud Machine Learning Engine and Tensor Flow allow unique access to machine learning tools without the need for PhD-educated data scientists.

The following companies use Google's tools to solve the listed problem.

1. *Axa International.* This global insurance company uses machine learning to predict which drivers would be more likely to cause major accidents. The analysis provides prediction accuracy of 78 percent. This prediction is used to determine appropriate insurance premiums.
2. *Airbus Defense & Space.* Detecting clouds in satellite imagery was done manually for decades. Using machine learning, the process has been expedited by 40 percent, and the error rate has been reduced from 11 percent to 3 percent.
3. *Preventing overfishing globally.* A government agency previously monitored only small sample regions globally to find fishing violators. Now, using satellite AIS positioning, the agency can watch the entire ocean. Using machine learning, the agency can track all fishing vessels to find violators.
4. *Detecting credit card fraud in Japan.* SMFG, a Japanese financial services company, uses Google's machine learning (a deep learning application) to monitor fraud related to credit card use, with an 80–90 percent accuracy of detection. The detection generates an alarm for taking actions.
5. *Kewpie Food of Japan.* This company detected defective potato cubes manually using a slow and expensive process. Using Google AI tools enables it to automatically monitor video feeds and alert inspectors to remove defective potatoes.

Source: Condensed and compiled from Forrest (2017).

Questions for Case 2.3

1. Why use machine learning for predictions?
2. Why use machine learning for detections?
3. What specific decisions were supported in the five cases?

INTELLIGENT AND AUTOMATED DECISION SUPPORT As early as 1970, there were attempts to automate decision making. These attempts were typically done with the use of rule-based expert systems that provided recommended solutions to repetitive managerial problems. Examples of decisions made automatically include the following:

- Small loan approvals
- Initial screening of job applicants
- Simple restocking
- Prices of products and services (when and how to change them)
- Product recommendation (e.g., at **Amazon.com**)

The process of automated decision making is illustrated in Figure 2.5. The process starts with knowledge acquisition and creation of a knowledge repository. Users submit questions to the system brain, which generates a response and submits it to the users. In addition, the solutions are evaluated so that the knowledge repository and the reasoning from it can be improved. Complex situations are forwarded to humans' attention. This process is especially used in knowledge-based systems. Note that the process in Figure 2.5 for knowledge acquisition illustrates automatic decision making as well. Companies use automated decision making for both their external operations (e.g., sales) and internal operations (e.g., resource allocation, inventory management). An example follows.

Example: Supporting Nurses' Diagnosis Decisions

A study conducted in a Taiwanese hospital (Liao, et al., 2015) investigated the use of AI to generate nursing diagnoses and compared them to diagnoses generated by humans. Diagnoses required comprehensive knowledge, clinical experience, and instinct. The researchers used several AI tools, including machine learning, to conduct data mining and analysis to predict the probable success of automated nursing diagnoses based on patient characteristics. The results indicated an 87 percent agreement between the AI and human diagnosis decisions.

Such technology can be used in places that have no human nursing staff as well as by nursing staff who want to verify the accuracy of their own diagnostic predictions. The system can facilitate the training of nursing staff as well.

Automated decisions can take several forms, as illustrated in Technology Insight 2.2.

Conclusion

There is little doubt that AI can change the decision-making process for businesses; for an example, see Sincavage (2017). The nature of the change varies based on the circumstances. But, in general, we expect AI to have a major impact for making better, faster, and more efficient decisions. Note that, in some cases, an AI watchdog is needed to regulate the process (see Sample, 2017, for details).

SECTION 2.5 REVIEW QUESTIONS

1. Distinguish between fully automated and supported decision making.
2. List the benefits of AI for decision support.
3. What factors influence the use of AI for decision support?
4. Relate AI to the steps in the classical decision-making process.
5. What are the necessary conditions for AI to be able to automate decision making?
6. Describe Schrage's four models.

TECHNOLOGY INSIGHT 2.2 Schrage's Models for Using AI to Make Decisions

Schrage (2017) of MIT's Sloan School has proposed the following four models for AI to make autonomous business decisions:

1. *The Autonomous Advisor.* This is a data-driven management model that uses AI algorithms to generate best strategies and instructions on what to do and makes specific recommendations. However, only humans can approve the recommendations (e.g., proposed solutions).

   Schrage provided an example in which an American retailing company replaced an entire merchandising department with an AI machine, ordering employees to obey directives from it. Obviously, resistance and resentment followed. To ensure compliance, the company had to install monitoring and auditing software.
2. *The Autonomous Outsource.* Here, the traditional business process outsourcing model is changed to a business process algorithm. To automate this activity, it is necessary to create crystal-clear rules and instructions. It is a complex scenario since it involves resource allocation. Correct predictability and reliability are essential.
3. *People–Machine Collaboration.* Assuming that algorithms can generate optimal decisions in this model, humans need to collaborate with the brilliant, but constrained, fully automated machines. To ensure such collaboration, it is necessary to train people to work with the AI machines (see the discussion in Chapter 14). This model is used by tech giants such as Netflix, Alibaba, and Google.
4. *Complete Machine Autonomy.* In this model, organizations fully automate entire processes. Management needs to completely trust AI models, a process that may take years. Schrage provides an example of a hedge fund that trades very frequently based on a machine's recommendations. The company uses machine learning to train the trading algorithms.

Implementing these four models requires appropriate management leadership and collaboration with data scientists. For suggestions of how to do so, consult Schrage (2017), who has written several related books. Kiron (2017) discusses why managers should consider AI for decision support.

An interesting note is that some competition among companies will actually occur among data-driven autonomous algorithms and related business models.

Questions for Discussion

1. Differentiate between the autonomous advisor and the people–machine collaboration models.
2. In all four models, there are some degrees of people–machine interaction. Discuss.
3. Why it is easier to use model 4 for investment decisions than, for example, marketing strategies?
4. Why is it important for data scientists to work with top management in autonomous AI machines?

2.6 AI APPLICATIONS IN ACCOUNTING

Throughout this book, we provide many examples of AI applications in business, services, and government. In the following five sections, we provide additional applications in the traditional areas of business: accounting; finance; human resource management; marketing, advertising, and CRM; and production-operation management.

AI in Accounting: An Overview

The CEO of SlickPie Accounting Software for small businesses, Chandi (2017), noticed trends among professional accountants: their use of AI, including bots in professional

routines, increased. Chandi observed that the major drivers for this are perceived savings in time and money and increased accuracy and productivity. The adoption has been rapid and it has been followed by significant improvements. An example is the execution of compliance procedures, where, for instance, Ernst & Young (EY) is using machine learning for detecting anomalous data (e.g., fraudulent invoices).

AI in Big Accounting Companies

Major users of AI are the big tax and accounting companies as illustrated in Application Case 2.4.

Accounting Applications in Small Firms

Small accounting firms also use AI. For example, Crowe Horwath of Chicago is using AI to solve complex billing problems in the healthcare industry. This helps its clients to deal with claims processing and reimbursements. The firm can now solve difficult problems that had previously resisted solutions. Many other applications are used with the support of AI, ranging from analyzing real estate contracts to risk analysis. It is only a question of time before even smaller firms will be able to utilize AI as well.

Application Case 2.4 How EY, Deloitte, and PwC Are Using AI

The big accounting companies use AI to replace or support human activities in tasks such as tax preparation, auditing, strategy consulting, and accountancy services. They mostly use NLP, robotic process automation, text mining, and machine learning. However, they use different strategies as described by Zhou (2017):

- EY attempts to show quick, positive return on investment (ROI) on a small scale. The strategy concentrates on business value. EY uses AI, for example, to review legal documents related to leasing (e.g., to meet new government regulations).
- PricewaterhouseCoopers (PwC) favors small projects that can be completely functioning in four weeks. The objective is to demonstrate the value of AI to client companies. Once demonstrated to clients, the projects are refined. PwC demonstrates 70 to 80 such projects annually.
- Deloitte Touche Tohmatsu Limited, commonly referred to as Deloitte, builds cases that guide AI-based projects for both clients and internal use. The objective is to facilitate innovation. One successful area is the use of NLP for review of large contracts that may include hundreds of thousands of legal documents. The company reduced such review time from six months to less than a month, and it reduced the number of employees who had performed the review by more than 70 percent. Deloitte, like its competitors, is using AI to evaluate potential procurement synergies for merger and acquisition decisions. Such evaluation is a time-consuming task since it is necessary to check huge quantities of data (sometime millions of data lines). As a result, Deloitte can finish such evaluation in a week compared to the four to five months required earlier. Deloitte said that with AI, it is viewing data in ways never even contemplated before (Ovaska-Few, 2017).

All big accounting companies use AI to assist in generating reports and to conduct many other routine, high-volume tasks. AI has produced high-quality work, and its accuracy has become better and better with time.

Sources: Compiled from Chandi (2017), Zhou (2017), and Ovaska-Few (2017).

Questions for Case 2.4

1. What are the characteristics of the tasks for which AI is used?
2. Why do the big accounting firms use different implementation strategies?

COMPREHENSIVE STUDY OF AI USE IN ACCOUNTING The ICAEW information technology (IT) faculty provides a free comprehensive study, "AI and the Future of Accountancy." This report (ICAEW, 2017) provides an assessment of AI use in accounting today and in the future. The report sees the advantage of AI by:

- Providing cheaper and better data to support decision making and solve accounting problems
- Generating insight from data analysis
- Freeing time of accountants to concentrate on problem solving and decision making

The report points to the use of the following:

- Machine learning for detecting fraud and predicting fraudulent activities
- Machine-learning and knowledge-based systems for verifying of accounting tasks
- Deep learning to analyze unstructured data, such as in contracts and e-mails

Job of Accountants

AI and analytics will automate many routine tasks done today by accountants (see discussion in Chapter 14), many of whom may lose their jobs. On the other hand, accountants will need to manage AI-based accounting systems. Finally, accountants need to drive AI innovation in order to succeed or even survive (see Warawa, 2017).

SECTION 2.6 REVIEW QUESTIONS

1. What are the major reasons for using AI in accounting?
2. List some applications big accounting firms use.
3. Why do big accounting firms lead the use of applied AI?
4. What are some of the advantages of using AI cited by the ICAEW report?
5. How may the job of the accountant be impacted by AI?

2.7 AI APPLICATIONS IN FINANCIAL SERVICES

Financial services are much diversified, and so is AI usage in the area. One way to organize the AI activities is by major segments of services. In this section, we discuss only two segments: banking and insurance.

AI Activities in Financial Services

Singh (2017) observed the following activities that may be found across various types of financial services:

- Extreme personalization (e.g., using chatbots, personal assistants, and robo investment advisors) (Chapter 12)
- Shifting customer behavior both online and in brick-and-mortar branches
- Facilitating trust in digital identity
- Revolutionizing payments
- Sharing economic activities (e.g., person-to-person loans)
- Offering financial services 24/7 and globally (connecting the world)

AI in Banking: An Overview

Consultancy.uk (2017) provides an overview of how AI is transforming the banking industry. It found AI applications mostly in IT, finance and accounting, marketing and sales, human resource management (HRM), customer service, and operations. A comprehensive

survey on AI in banking was conducted in 2017, and a report is available for purchase (see Tiwan, 2017).

The key findings of this report are as follows:

- AI technologies in banking include all those listed in Section 2.7 and several other analytical tools (Chapters 3 to 11 of this book).
- These technologies help banks improve both their front-office and back-office operations.
- Major activities are the use of chatbots to improve customer service and communicating with customers (see Chapter 12), and robo advising is used by some financial institutions (see Chapter 12).
- Facial recognition is used for safer online banking.
- Advanced analytics helps customers with investment decisions. For examples of this help, see Nordrum (2017), E. V. Staff (2017), and Agrawal (2018).
- AI algorithms help banks identify and block fraudulent activities including money laundering.
- AI algorithms can help in assessing the creditworthiness of loan applicants. (For a case study of an application of AI in credit screening, see **ai-toolkit.blogspot.com/2017/01/case-study-artificial-intelligence-in.html**.)

Illustrative AI Applications in Banking

The following are banking institutions that use AI:

- Banks are using AI machines, such as IBM Watson, to step up employee surveillance. This is important in preventing illegal activities such as those that occurred at Wells Fargo, the financial services and banking company. For details, see **information-management.com/articles/banks-using-algorithms-to-step-up-employee-surveillance**.
- Banks use applications for tax preparation. H&R Block is using IBM Watson to review tax returns. The program makes sure that individuals pay only what they owe. Using interactive conversations, the machine attempts to lower people's tax bills.
- Answering many queries in real time. For example, Rainbird Co. (**rainbird.ai/**) is an AI vendor that trains machines to answer customers' queries. Millions of customers' questions keep bank employees busy. Bots assist staff members to quickly find the appropriate answers to queries. This is especially important in banks where turnovers of employees are high. Also, there is knowledge degrading overtime, due to frequent changes in policies and regulations.

  Rainbird is integrated with IBM Watson, which is using AI capabilities and cognitive reasoning to understand the nature of the queries and provide solutions. The program–employee conversations are done via chatbots, which are deployed to all U.K. branches of the banks served by Rainbird.
- At Capital One and several other banks, customers can talk with Amazon's Alexa to pay credit card bills and check their accounts.
- TD Bank and others (see Yurcan, 2017) experiment with Alexa, which provides machine learning and augmented reality capabilities for answering queries.
- Bank Danamon uses machine learning for fraud detection and anti–money-laundering activities. It also improves the customer experience.
- At HSBC, customers can converse with the virtual banking assistant, Olivia, to find information about their accounts and even learn about security. Olivia can learn from its experiences and become more useful.

- Santander Bank employs a virtual assistant (called Nina) that can transfer money, pay bills, and do more. Nina can also authenticate its customers via an AI-based voice recognition system. Luvo of RBS is a customer service and customer relationship management (CRM) bot that answers customers' queries.
- At Accenture, Collette is a virtual mortgage advisor that provides personalized advice.
- A robot named NaO can analyze facial expression and behavior of customers that enter the branches of certain banks and determine their nationality. Then the machine selects a matching language (Japanese, Chinese, or English) to interact with the customer.

IBM Watson can provide banks many other services ranging from crime fighting to regulatory compliance as illustrated next.

Example: How Watson Helps Banks Manage Compliance and Supports Decision Making

Government regulations place a burden on banks and other financial institutions. To comply with regulations, banks must spend a considerable amount of time examining huge amounts of data generated daily.

Developed by Promontory Financial Group (an IBM subsidiary), IBM Watson (Chapter 6) developed a set of tools to deal with the compliance problem. The set of tools was trained by using the knowledge of former regulators and examining data from over 200 different sources. All in all, the program is based on over 60,000 regulatory citations. It includes three sets of cognitive tools that deal with regulatory compliance. One of the tools deals with financial crimes, flagging potential suspicious transactions and possible fraud. The second tool monitors compliance, and the third one deals with the large volume of data. Watson is acting as a banking financial consultant for these and other banking issues.

IBM's tools are designed to assist financial institutions to justify important decisions. The AI algorithms examine the data inputs and outputs in managerial decision making. For example, when the program spots suspicious activity, it will notify the appropriate manager, who then will take the necessary action. For details, see Clozel (2017).

Application Case 2.5 illustrates US Bank's use of AI to improve customer service.

Insurance Services

Advancements in AI are improving several areas in the insurance industry, mostly in issuing policies and handling claims.

According to Hauari (2017), the major objectives of the AI support are to improve analysis results and enhance customer experience. Incoming claims are analyzed by AI, and, depending on their nature, are sent to appropriate available adjusters. The technologies used are NLP and text recognition (Chapters 6 and 7). The AI software can help in data collection and analysis and in data mining old claims.

Agents previously spent considerable time asking routine questions from people submitting insurance claims. AI machines, according to Beauchamp (2016), provide speed, accuracy, and efficiency in performing this process. Then AI can facilitate the underwriting process.

Similarly, claims processing is streamlined with the help of AI. It reduces processing time (by up to 90 percent) and improves accuracy. Capabilities of machine-learning and other AI programs can be shared in seconds in multi-office configurations, including global settings.

Application Case 2.5 US Bank Customer Recognition and Services

As of July 2017, US Bank has been able to automatically identify military service members and veterans when they call or enter one of its branches This is not a simple task. The service members are recognized by Einstein, an AI-based CRM service from Salesforce Inc. (see Section 2.9).

What US Bank is trying to do is to recognize customers and understand their needs. Einstein helps the bank gain a competitive advantage in doing so. Knowledge provided is important not only for marketing and providing targeted professional financial services but also for greeting customers on their birthdays or thanking them for using the bank's services.

The bank now has considerable information about customers available to human agents in real time. Such information helps customers when online and when at one of the bank's actual locations.

The AI application tells the rep all about the customer so the rep can offer appropriate services. For example, if the customer needs insurance, the AI will detect this need and the rep will offer a good alternative. It also offers information to an online customer: "Hello, Mary; I see you are checking your mortgage payments. I have good news for you. . . ."

Source: Compiled from Crosman (2017) and Carey (2017).

Questions for Case 2.5

1. What are Einstein's advantages to US Bank?
2. What are its advantages to customers?
3. What are the benefits of voice communication?

Insurers, like other adopters of AI, will have to go through a transformation and adapt to change. Companies and individual agents can learn from early adopters. For how this is done at MetLife, see Blog (2017).

Example: Metromile Uses AI in Claim Processing

Metromile is an innovator in vehicle insurance, using the pay-per-mile model. It operates in seven U.S. states. In mid-2017, it started using AI-based programs to automate accident data, process accident claims, and pay customer claims. The automated platform, according to Santana (2017), is powered by a smart claim bot called AVA. It processes images forwarded by customers, extracting the pertinent telematic data. The AI bot simulates the accidents' major points and makes a verification based on decision rules; authorization for payments provides for successful verification. The process takes minutes. Only complex cases are sent to investigation by human processors. Customers are delighted since they can get fast resolutions. While at the moment AVA is limited to certain types of claims, its range of suitability is increasing with the learning capabilities of machine learning and the advances in AI algorithms.

Note: A 2015 start-up, Lemonade (**lemonade.com**) provides an AI-based platform for insurance that includes bots and machine learning. For details, see Gagliordi (2017).

SECTION 2.7 REVIEW QUESTIONS

1. What are the new ways that banks interact with customers by using AI?
2. It is said that financial services are more personalized with AI support. Explain.
3. What back-office activities in banks are facilitated by AI?
4. How can AI contribute to security and safety?

5. What is the role of chatbots and virtual assistants in financial services?
6. How can IBM Watson help banking services?
7. Relate Salesforce Einstein to CRM in financial services.
8. How can AI help in processing insurance claims?

2.8 AI IN HUMAN RESOURCE MANAGEMENT (HRM)

As in other business functional areas, the use of AI technologies is spreading rapidly in HRM. And as in other areas, the AI services reduce cost and increase productivity, consistency, and speed of execution.

AI in HRM: An Overview

Savar (2017) points to the following reasons for AI to transform HRM, especially in recruiting: (1) reducing human bias, (2) increasing efficiency, productivity, and insight in evaluating candidates, and (3) improving relationships with current employees.

Wislow (2017) sees the use of AI as a continuation of automation that supports HRM and keeps changing it. Wislow suggests that such automation changes how HRM employees work and are engaged. This change also strengthens teamwork. Wislow divided the impact of AI into the following areas:

RECRUITMENT (TALENT ACQUISITION) One of the cumbersome tasks in HRM, especially in large organizations, is recruiting new employees. The fact is that many job positions are unfilled due to difficulties in finding the right employees. At the same time, many qualified people cannot find the right jobs.

AI improves the recruiting process as illustrated in Application Case 2.6.

The use of chatbots to facilitate recruitment is also described by Meister (2017).

Companies that help recruiters and job seekers, especially LinkedIn, are using AI algorithms to suggest matches to both recruiters and job seekers. Haines (2017) describes the process, noting that a key benefit of this process is the removal of unconscious biases and prejudices of humans.

AI FACILITIES TRAINING The rapid technological developments make it necessary to train and retrain employees. AI methods can be used to facilitate learning. For example, chatbots can be used as a source of knowledge to answer learners' queries. Online courses are popular with employees. AI can be used to test progress, for example. In addition, AI can be used to personalize online teaching for individuals and to design group lectures.

AI SUPPORTS PERFORMANCE ANALYSIS (EVALUATION) AI tools enable HR management to conduct performance analysis by breaking work into many small components and by measuring the performance of each employee and team on each component. The performance is compared to objectives, which are provided to employees and teams. AI also can track changes and progress by combining AI with analytical tools.

AI USE IN RETENTION AND ATTRITION DETECTION In order to keep employees from leaving, it is necessary for businesses to analyze and predict how to make workers happy. Machine learning can be used to detect reasons why employees leave companies by identifying influencing patterns.

AI in Onboarding

Once new employees are hired, the HR department needs help introducing them to the organizational culture and operating processes. Some new employees require much

Application Case 2.6

How Alexander Mann Solutions (AMS) Is Using AI to Support the Recruiting Process

Alexander Mann is a Chicago-based company that offers AI solutions to support the employee recruitment process. The major objective is to help companies solve HRM problems and challenges. The AI is used to:

1. Help companies evaluate applicants and their resumes by using machine learning. The result is the decision regarding which applicants to invite for an interview.
2. Help companies evaluate resumes that are posted on the Web. The AI software can use key words for the search related to the background of employees (e.g., training, years of experience).
3. Evaluate the resumes of the best employees who currently work in a company and create, accordingly, desired profiles to be used when vacancies occur. These profiles are then compared to those of applying candidates, and the top ones are ranked by their fit to each job opening. In addition to the ranking, the AI program shows the fit with each desired criterion. At this stage, the human recruiter can make a final selection decision. This way, the selection process is faster and much more accurate.

The accuracy of the process solves the candidate volume problem, ensuring that qualified people are not missed and poorly fit applicants are not selected.

Alexander Mann is also helping its clients to install chatbots that can provide candidates' answers to questions related to the jobs and the working conditions at the employing company. (For the recruiting chatbot, see Dickson, 2017).

Sources: Compiled from Huang (2017), Dickson (2017), and **alexandermannsolutions.com**, accessed June 2018.

Questions for Case 2.6

1. What types of decisions are supported?
2. Comment on the human–machine collaboration.
3. What are the benefits to recruiters? To applicants?
4. Which tasks in the recruiting process are fully automated?
5. What are the benefits of such automation?

attention. AI helps HRM prepare customized onboarding paths that are best for the newcomers. Results showed that those employees supported by AI-based plans tend to stay longer in organizations (Wislow, 2017).

USING CHATBOTS FOR SUPPORTING HRM The use of chatbots in HRM is increasing rapidly. Their ability to provide current information to employees anytime is a major reason. Dickson (2017) refers to the following chatbots: Mya, a recruiting assistant, and Job Bot, which supports the recruitment of hourly workers. This bot is also used as a plug-in to Craigslist. Another chatbot mentioned earlier is Olivia; see **olivia.paradox.ai/**.

Introducing AI to HRM Operations

Introducing AI to HRM operations is similar to introducing AI to other functional areas.

Meister (2017) suggests the following activities:

1. Experiment with a variety of chatbots
2. Develop a team approach involving other functional areas
3. Properly plan a technology roadmap for both the short and long term, including shared vision with other functional areas
4. Identify new job roles and modifications in existing job roles in the transformed environment
5. Train and educate the HRM team to understand AI and gain expertise in it

For additional information and discussion, see Essex (2017).

SECTION 2.8 REVIEW QUESTIONS

1. List the activities in recruiting and explain the support provided by AI to each.
2. What are the benefits rewarded to recruiters by AI?
3. What are the benefits to job seekers?
4. How does AI facilitate training?
5. How is performance evaluation of employees improved by AI?
6. How can companies increase retention and reduce attrition with AI?
7. Describe the role of chatbots in supporting HRM.

2.9 AI IN MARKETING, ADVERTISING, AND CRM

Compared to other business areas, there are probably more applications of AI in marketing and advertising. For example, AI-based product recommendations have been in use by **Amazon.com** and other e-commerce companies for more than 20 years. Due to the large number of applications, we provide only a few examples here.

Overview of Major Applications

Davis (2016) provides 15 examples of AI in marketing as listed with explanations by the authors of this book and from Martin (2017). Also see Pennington (2018).

1. *Product and personal recommendations.* Starting with **Amazon.com**'s book recommendations for Netflix's movies, AI-based technologies are used extensively for personalized recommendations (e.g., see Martin, 2017).
2. *Smart search engines.* Google is using RankBrain's AI system to interpret users' queries. Using NLP helps in understanding the products or services for which online users are searching. This includes the use of voice communication.
3. *Fraud and data breaches detection.* Application for this has covered credit/debit card use for many years, protecting Visa and other card issuers. Similar technologies protect retailers (such as Target and Neiman Marcus) from hackers' attacks.
4. *Social semantics.* Using AI-based technologies, such as sentiment analysis and image and voice recognitions, retailers can learn about customers' needs and provide targeted advertisements and product recommendations directly (e.g., via e-mail) and through social media.
5. *Web site design.* Using AI methods, marketers are able to design attractive Web sites.
6. *Producer pricing.* AI algorithms help retailers price products and services in a dynamic fashion based on the competition, customers' requirements, and more. For example, AI provides predictive analysis to forecast the impact of different price levels.
7. *Predictive customer service.* Similar to predicting the impact of pricing, AI can help in predicting the impact of different customer service options.
8. *Ad targeting.* Similar to product recommendations, which are based on user profiles, marketers can tailor ads to individual customers. The AI machines attempt to match different ads with individuals.
9. *Speech recognition.* As the trend to use voice in human–machine interaction is increasing, the use of bots by marketers to provide product information and prices accelerates. Customers prefer to talk to bots rather than to key in dialogue.
10. *Language translation.* AI enables conversations between people who speak different languages. Also, customers can buy from Web sites written in languages they do not speak by using GoogleTranslate, for example.

11. *Customer segmentation.* Marketers are segmenting customers into groups and then tailoring ads to each group. While less effective than targeting individuals, this is more effective than mass advertising. AI can use data and text mining to help marketers identify the characteristics of specific segments (e.g., by mining historical files) as well as help tailor the best ads for each segment.
12. *Sales forecasting.* Marketers' strategy and planning are based on sales forecasting. Such forecasting may be very difficult for certain products. Uncertainties may exist in many situations such as in customer need assessment. Predictive analytics and other AI tools can provide better forecasting than traditional statistical tools.
13. *Image recognition.* This can be useful in market research (e.g., for identifying consumer preferences of one company's products versus those of its competition). It can also be used for detecting defects in producing and/or packaging products.
14. *Content generation.* Marketers continuously create ads and product information. AI can expedite this task and make sure that it is consistent and complies with regulations. Also, AI can help in generating targeted content to both individuals and segments of consumers.
15. *Using bots, assistants, and robo advisors.* In Chapter 12, we describe how bots, personal assistants, and robo advisors help consumers of products and services. Also, these AI machines excel in facilitating customer experience and strengthen customer relationship management. Some experts call bots and virtual personal assistants the "face of marketing."

Another list is provided at **en.wikipedia.org/wiki/Marketing_and_artificial_intelligence**.

AI Marketing Assistants in Action

There are many ways that AI can be used in marketing. One way is illustrated in Application Case 2.7 about Kraft Foods.

Customer Experiences and CRM

As described earlier, a major impact of AI technologies is changing customer experiences. A notable example is the use of conversational bots. Bots (e.g., Alexa) can provide information about products and companies and can provide advice and guidance (e.g., robo advisors for investment; see Chapter 12). Gangwani (2016) lists the following ways to improve customers' experiences:

1. Use NLP for generating user documentation. This capability also improves the customer–machine dialogue.
2. Use visual categorization to organize images (for example, see IBM's Visual Recognition and Clarifai)
3. Provide personalized and segmented services by analyzing customer data. This includes improving shopping experience and CRM.

A well-known example of AI in CRM is Salesforce's Einstein.

Example: Salesforce's AI Einstein

Salesforce Einstein is an AI set of technologies (e.g., Einstein Vision for image recognition) that is used for enhancing customer interactions and supporting sales. For example, the system delivers dynamic sales dashboards to sales reps. It also tracks performance and manages teamwork by using sales analytics. The AI product also can provide predictions

Application Case 2.7 Kraft Foods Uses AI for Marketing and CRM

The number of mobile users is growing rapidly as is the number of mobile shoppers. Kraft Foods took notice of that. The company is adapting its advertising and sales to this trend. Mobile customers are looking for brands and interacting with Kraft brands. Kraft Foods wanted to make it easy for customers to interact with the company whenever and wherever they want. To achieve this interaction goal, Kraft Foods created a "Food Assistant," also known as Kraft Food Assistant.

The Kraft Food Assistant

Kraft's Food Assistant is an app for smartphones that allows customers to access more than 700 recipes. Thus, the consumer can browse easily for ideas. Customers enter a virtual store and open the "recipe of the day." The app tells the user all the ingredients needed for that recipe or for any desired recipe. The Food Assistant also posts all the relevant coupons available for the ingredients on users' smartphone. Users need only to take the smartphone to a supermarket, scan the coupons, and save on the ingredients. The recipe of the day is also demonstrated on a video. Unique to this app is the inclusion of an AI algorithm that learns from users' orders and can infer, for example, the users' family size. The more the AI learns about users, the more suggestions it makes. For example, it tells users what to do with their leftover ingredients. In addition, the more the Food Assistant learns about users, the more useful suggestions for recipes and cooking it can offer. It is like the Netflix recommender. The more Kraft products that users buy (the ingredients), the more advice they get. The Food Assistant also directs users to the nearest store that has the recipes' ingredients. Users can get assistance on how to prepare food in 20 minutes and on many cooking-related topics.

The AI is tracking consumers' behavior. Information is stored on each user's loyalty card. The system makes inferences about what consumers like and targets related promotions to them. This process is called *behavioral pattern recognition,* and is based on AI techniques such as "collaborative filtering." (See Chapter 12.)

AI assistants also can tweak messages to users, and they know if users are interested in their topics. The assistant also knows whether customers are responding positively and whether they are or are not motivated to try a new product or purchase more of what they previously purchased. The Kraft AI Food Assistant actually is trying to *influence* and sometimes to *modify* consumer behavior. Like other vendors, Kraft is using the information collected by the AI assistant to forge and execute mobile and regular commerce strategies.

Using the information collected, Kraft and similar vendors can expand their mobile marketing programs both online and in physical stores.

Note: Users can interact with the system with voice powered by Nuance Communication. The system is based on natural language processing.

Sources: Compiled from Celentano (2016), press releases at **nuance.com**, and **kraftrecipes.com/media/iphoneassistant.aspx/**, accessed March 2018.

Questions for Case 2.7

1. Identify all AI technologies used in the Food Assistant.
2. List the benefits to the customers.
3. List the benefits to Kraft Foods.
4. How is advertising done?
5. What role is "behavioral pattern recognition" playing?
6. Compare Kraft's Food Assistant to **Amazon.com** and Netflix recommendation systems.

and recommendations. It supports Salesforce Customer Successful Platform and other Salesforce products.

Einstein's automatically prioritized sales leads make sales reps more productive when dealing with sales leads and potential opportunities. The sales reps also get insights about customers' sentiments, competitors' involvement, and other information.

For information and a demo, see **salesforce.com/products/einstein/overview/**. For features and description of the product, see **zdnet.com/article/salesforces-einstein-ai-platform-what-you-need-to-know/**. For additional features, see **salesforce.com/products/einstein/features/**.

Other Uses of AI in Marketing

The following show the diversity of AI technologies used in marketing:

- It is used to mimic the expertise of in-store salespeople. In many physical stores, humans are not readily available to help customers who do not want to wait very long. Thus, shopping is made easier when bots provide guidance. A Japanese store already provides all services in a physical store by speaking robots.
- It provides lead generation. As seen in the case of Einstein, AI can help generate sales leads by analyzing customers' data. The program can generate predictions. Insights can be generated by intelligent analytics.
- It can increase customer loyalty using personalization. For example, some AI techniques can recognize regular customers (e.g., in banks). IBM Watson can learn about people from their tweets.
- **Salesforce.com** provides a free e-book, "Everything You Need to Know about AI for CRM" (**salesforce.com/form/pdf/ai-for-crm.jsp**).
- It can improve the sales pipeline. Narayan (2018) provides a process of how companies can use AI and robots to do this. Specifically, robots convert unknown visitors into customers. Robots use three stages: (1) prepare a list of target customers in the database, (2) send information, ads, videos, and so on to prospects on the list created earlier, and (3) provide the company sales department with a list of leads that successfully convert potential customers to buyers.

SECTION 2.9 REVIEW QUESTIONS

1. List five of the 15 applications of Davis (2016). Comment on each.
2. Which of the 15 applications relate to sales?
3. Which of the 15 applications relate to advertising?
4. Which of the 15 applications relate to customer service and CRM?
5. For what are the prediction capabilities of AI used?
6. What is the Salesforce's Einstein?
7. How can AI be used to improve CRM?

2.10 AI APPLICATIONS IN PRODUCTION-OPERATION MANAGEMENT (POM)

The field of POM is much diversified, and its use of AI is evident today in many areas. To describe all of them, we would need more than a whole book. In the remaining chapters, we provide dozens of examples about AI applications in POM. Here, we provide only a brief discussion regarding two related application areas: manufacturing and logistics.

AI in Manufacturing

To handle ever-increasing labor costs, changes in customers' requirements, increased global competition, and government regulations (Chapter 1), manufacturing companies are using elevated levels of automation and digitization. According to Bollard et al.

(2017), companies need to be more agile, and react quicker and more effectively. They also need to be more efficient and improve customers' (organizations' and individuals') experiences. Companies are pressured to cut costs and increase quality and transparency. To achieve these goals, they need to automate processes and make use of AI and other cutting-edge technologies.

Implementation Model

Bollard, et al. (2017) proposed a five-component model for manufacturing companies to use intelligent technologies. This model includes:

- Streamlining processes, including minimizing waste, redesigning processes, and using business process management (BPM)
- Outsourcing certain business processes, including going offshore
- Using intelligence in decision making by deploying AI and analytics
- Replacing human tasks with intelligent automation
- Digitizing customers' experiences

Companies have used this model for a long time. Actually, robotics have been used since around 1960 (e.g., Unimate in General Motors). However, the robots were "dumb," each usually doing one simple task. Today, companies use intelligent robots for complex tasks, enabling make-to-order products and mass customization. In other words, many mental and cognitive tasks are being automated. These developments, involving AI and sensors, allow supporting or even automating production decisions in real time.

Example

When a sensor detects a defective product or a malfunction, the data are processed by an AI algorithm. An action then takes place instantly and automatically. For example, a defective item can be removed or replaced. AI can even make predictions about equipment failures before they occur (see the opening vignette in Chapter 1). This real-time action saves a huge amount of money for manufacturers. (This process may involve the IoT; see Chapter 13.)

Intelligent Factories

Ultimately, companies will use smart or intelligent factories (see Chapter 13). These factories use complex software and sensors. An example of a lead supplier is General Electric, which provides software such as OEE Performance Analyzer and Production Execution Supervisor. The software is maintained in the "cloud" and it is provided as a "software-as-a-service." GE partners with Cisco and FTC to provide security, connectivity, and special analytics.

In addition to GE, well-known companies such as Siemens and Hitachi provide comprehensive solutions. For an example, see Hitachi AI Technology's Report (**social-innovation.hitachi/ph/solutions/ai/pdf/ai_en_170310.pdf**).

Many small vendors are specializing in different aspects of AI for manufacturing. For example, BellHawk Systems Corporation, which provides services to small companies, specializes in real-time operations tracking (see Green, 2016).

Early successes were recorded by large companies such as Procter & Gamble and Toyota.

However, as time passes, medium-size and small companies can also afford AI services. For additional information, see **bellhawk.com**.

Logistics and Transportation

AI and intelligent robots are used extensively in corporate logistics and internal and external transportation, as well as in supply chain management. For example, **Amazon.com** is using over 50,000 robots to move items in its distribution centers (other e-commerce companies are doing the same). Soon, we will see driverless trucks and other autonomous vehicles all over the world (see Chapter 13).

Example: DHL Supply Chain

DHL is a global delivery company (competing with FedEx and UPS). It has a supply chain division that works with many business partners. AI and IoT are changing the manner by which the company, its partners, and even its competitors operate. DHL is developing innovative logistics and transportation business models, mostly with AI, IoT, and machine learning. These models also help DHL's customers gain a competitive advantage (and this is why the company cannot provide details in its reports).

Several of the IoT projects are linked to machine learning, specifically in the areas of sensors, communication, device management, security, and analysis. Machine learning in such cases assists in tailoring solutions to specific requirements.

Overall, DHL concentrates on the areas of supply chains (e.g., identifies inventories and controls them along the supply chain) and warehouse management. Machine learning and other AI algorithms enable more accurate procurement, production planning, and work coordination. Tagging and tracking items using Radio Frequency Identification (RFID) and Quick Response (QR) code allow for item tracking along the supply chain. Finally, AI facilitates predictive analytics, scheduling, and resource planning. For details, see Coward (2017).

SECTION 2.10 REVIEW QUESTIONS

1. Describe the role of robots in manufacturing.
2. Why use AI in manufacturing?
3. Describe the Bollard et al. implementation model.
4. What is an intelligent factory?
5. How are a company's internal and external logistics supported by AI technologies?

Chapter Highlights

- The aim of artificial intelligence is to make machines perform tasks intelligently, possibly like people do.
- A major reason for using AI is to cause work and decision making to be easier to perform. AI can be more capable (enable new applications and business models), more intuitive, and less threatening than other decision support applications.
- A major reason to use AI is to reduce cost and/or increase productivity.
- AI systems can work autonomously, saving time and money, and perform work consistently. They can also work in rural and remote areas where human expertise is rare or not available.
- AI can be used to improve all decision-making steps.
- Intelligent virtual systems can act as assistants to humans.
- AI systems are computer systems that exhibit low (but increasing) levels of intelligence.
- AI has several definitions and derivatives, and its importance is growing rapidly. The U.S. government postulated that AI will be a "critical driver of the U.S. economy" (Gaudin 2016).
- The major technologies of AI are intelligent agents, machine learning, robotic systems, NLP and speech recognition, computer vision, and knowledge systems.

- Expert systems, recommendation systems, chatbots, and robo advisors are all based on knowledge transferred to machines.
- The major limitations of AI are the lack of human touch and feel, the fear that it will take jobs from people, and the possibility that it could be destructive.
- AI is not a match to humans in many cognitive tasks, but it can perform many manual tasks quicker and at a lower cost.
- There are several types of intelligence, so it is difficult to measure AI's capacity.
- In general, human intelligence is superior to that of machines. However, machines can beat people in complex games.
- Machine learning is currently the most useful AI technology. It attempts to learn from its experience to improve operations.
- Deep learning enables AI technologies to learn from each other, creating synergy in learning.
- Intelligent agents excel in performing simple tasks considerably faster and more consistently than humans (e.g., detecting viruses in computers).
- The major power of machine learning is a result of the machine's ability to learn from data and its manipulation.
- Deep learning can solve many difficult problems.
- Computer vision can provide understandings from images, including from videos.
- Robots are electromechanical computerized systems that can perform physical and mental tasks. When provided with sensory devices, they can become intelligent.
- Computers can understand human languages and can generate text or voice in human languages.
- Cognitive computing simulates the human thought process for solving problems and making decisions.
- Computers can be fully automated in simple manual and mental tasks using AI.
- Several types of decision making are fully automated using AI; other types can be supported.
- AI is used extensively in all functional business departments, reducing cost and increasing productivity, accuracy, and consistency. There is a tendency to increase the use of chatbots. They all support decision making well.
- AI is used extensively in accounting, automating simple transactions, helping deal with Big Data, finding fraudulent transactions, increasing security, and assisting in auditing and compliance.
- AI is used extensively in financial services to improve customer service, provide investment advice, increase security, and facilitate payments among other tasks. Notable applications are in banking and insurance.
- HRM is using AI to facilitate recruitment, enhance training, help onboarding, and streamline operations.
- There is considerable use of AI in marketing, sales, and advertising. AI is used to support product recommendation, help in search of products and services, facilitate Web site design, support pricing decisions, provide language translation in globe trade, assist in forecasting and predictions, and use chatbots for many marketing and customer service activities.
- AI has been used in manufacturing for decades. Now it is applied to support planning, supply chain coordination, logistics and transportation, and operation of intelligent factories.

Key Terms

artificial brain
artificial intelligence (AI)
augmented intelligence
chatbots
computer vision
deep learning
intelligent agent
machine learning
machine vision
natural language processing (NLP)
robot
scene recognition
shopbot
speech (voice) understanding
Turing Test

Questions for Discussion

1. Discuss the difficulties in measuring the intelligence of machines.
2. Discuss the process that generates the power of AI.
3. Discuss the differences between machine learning and deep learning.

4. Describe the difference between machine vision and computer vision.
5. How can a vacuum cleaner be as intelligent as a six-year-old child?
6. Why are NLP and machine vision so prevalent in industry?
7. Why are chatbots becoming very popular?
8. Discuss the advantages and disadvantages of the Turing Test.
9. Why is augmented reality related to AI?
10. Discuss the support that AI can provide to decision makers.
11. Discuss the benefits of automatic and autonomous decision making.
12. Why is general (strong) AI considered to be "the most significant technology ever created by humans"?
13. Why is the cost of labor increasing, whereas the cost of AI is declining?
14. If an artificial brain someday contains as many neurons as the human brain, will it be as smart as a human brain? (Students need to do extra research.)
15. Distinguish between dumb robots and intelligent ones.
16. Discuss why applications of natural language processing and computer vision are popular and have many uses.

Exercises

1. Go to **itunes.apple.com/us/app/public-transit-app-moovit/id498477945?mt=8**. Compare Moovit operations to the operation of INRIX.
2. Go to **sitezeus.com** and view the 2:07 min. video. Explain how the technology works as a decision helper.
3. Go to Investopedia and learn about investors' tolerance. Then find out how AI can be used to contain this risk, and write a report.
4. In 2017, McKinsey & Company created a five-part video titled "Ask the AI Experts: What Advice Would You Give to Executives About AI?" View the video and summarize the advice given to the major issues discussed. (Note: This is a class project.)
5. Watch the McKinsey & Company video (3:06 min.) on today's drivers of AI at **youtube.com/watch?v=yv0IG1D-OdU** and identify the major AI drivers. Write a report.
6. Go to the Web site of the Association for the Advancement of Artificial Intelligence **aaai.org/home.html** and describe its content. Compare it to that of **ai.sri.com** and **csail.mit.edu/**.
7. Go to **crosschx.com** and find information about Olive. Explain how it works, what its limitations and advantages are, and which types of decisions it automates and which it only supports.
8. Go to **waze.com** and **moovitapp.com** and find their capabilities. Summarize the help they can provide users.
9. Go to **sentient.ai**. Find its products that facilitate e-commerce. Write a report.
10. Go to **artificialbrain.org** and report the latest progress there.
11. Find recent information on research that is aimed to measure artificial intelligence. Write a report.
12. Go to **salesforce.com** and find recent developments on AI Einstein. Why it is so popular?
13. Find the latest information on IBM Watson's advising activities. Write a report.
14. Find information on the use of AI in iPhones. Explore the role of Edge AI. Write a report.
15. Explore the AI-related products and services of Nuance Inc. (**nuance.com**). Explore the Dragon voice recognition product.
16. Go to the Netradyne report at **cs_netradyne.com/** and read about the use of its product for road safety. Write a report.
17. Go to **salesforce.com** and investigate the capabilities of Gecko HRM. Relate it to Salesforce Einstein. Provide examples of two applications.
18. Enter McKinsey & Company and find in its Fifty Five "The Value AI Can Bring to Your Business" (**mckinsey.com/featured-insights/artificial-intelligence/five-fifty-real-world-ai**). Then look for "Real-World AI." Find the banking section and dive more deeply into its content.
19. Find material on the impact of AI on advertising. Write a report.
20. Go to **strategicsourceror.com/2018/03/giant-scale-supply-chains-can-make.html**. Summarize the use of AI.

References

Agrawal, V. "How Successful Investors Are Using AI to Stay Ahead of the Competition." *ValueWalk*, January 28, 2018.

Alpaydin, E. *Machine Learning: The New AI (The MIT Press Essential Knowledge Series)*. Boston, MA: MIT Press, 2016.

Beauchamp, P. "Artificial Intelligence and the Insurance Industry: What You Need to Know." *The Huffington Post*, October 27, 2016.

Blog. "Welcome to the Future: How AI Is Transforming Insurance." **Blog.metlife.com**, October 1, 2017.

Bollard, A., et al. "The next-generation operating model for the digital world." *McKinsey & Company*, March 2017.

BrandStudio. "Future-Proof: How Today's Artificial Intelligence Solutions Are Taking Government Services to the Next Frontier." *Washington Post*, August 22, 2017.

Carey, S. "US Bank Doubles Its Conversion Rate for Wealth Customers Using Salesforce Einstein." *Computerworld UK*, November 10, 2017.

Carney, P. "Pat Carney: Artificial Intelligence versus Human Intelligence." *Vancouver Sun*, April 7, 2018.

Celentano, D. "Kraft Foods iPhone Assistant Appeals to Time Starved Consumers." *The Balance*, September 18, 2016.

Chandi, N. "How AI is Reshaping the Accounting Industry." **Forbes.com**, July 20, 2017.

Clozel, L. "IBM Unveils New Watson tools to Help Banks Manage Compliance, AML." *American Banker*, June 14, 2017.

Consultancy.uk. "How Artificial Intelligence Is Transforming the Banking Industry." September 28, 2017. **consultancy.uk/news/14017/how-artificial-intelligence-is-transforming-the-banking-industry/** (accessed June 2018).

Coward, J. "Artificial Intelligence Is Unshackling DHL's Supply Chain Potential." *IoT Institute*, April 18, 2017. **ioti.com/industrial-iot/artificial-intelligence-unshackling-dhls-supply-chain-potential** (accessed June 2018).

Crosman, P. "U.S. Bank Bets AI Can Finally Deliver 360-Degree View." *American Banker*, July 20, 2017.

Davis, B. "15 Examples of Artificial Intelligence in Marketing." *Econsultancy*, April 19, 2016.

Dickson, B. "How Artificial Intelligence Optimizes Recruitment." *The Next Web*, June 3, 2017.

Dodge, J. "Artificial Intelligence in the Enterprise: It's On." *Computerworld*, February 10, 2016.

Dormehl, L. *Thinking Machines: The Quest for Artificial Intelligence—and Where It's Taking Us Next*. New York, NY: Tarcher-Perigee, 2017.

Essex, D. "AI in HR: Artificial Intelligence to Bring Out the Best in People." *TechTargetEssential Guide*, April 2017.

E. V. Staff. "Artificial Intelligence Used to Predict Short-Term Share Price Movements." *The Economic Voice*, June 22, 2017.

Finlay, S. *Artificial Intelligence and Machine Learning for Business: A No-Nonsense Guide to Data Driven Technologies*. 2nd ed. Seattle, WA: Relativistic, 2017.

Forrest, C. "7 Companies That Used Machine Learning to Solve Real Business Problems." *Tech Republic*, March 8, 2017.

Fuller, D. "LG Claims Its Roboking Vacuum Is As Smart As a Child." *Androidheadlines.com*, July 18, 2017.

Gagliordi, N. "Softbank Leads $120M Investment in AI-Based Insurance Startup Lemonade." *ZDNET*, December 19, 2017.

Gangwani, T. "3 Ways to Improve Customer Experience Using A.I." *CIO Contributor Network*, October 12, 2016.

Gaudin, S. "White House: A.I. Will Be Critical Driver of U.S. Economy." *Computerworld*, October 12, 2016.

Gitlin, J. M. "Watch Out, Waze: INRIX's New Traffic App Is Coming for You." *Ars Technica*, March 30, 2016. **arstechnica.com/cars/2016/watch-out-waze-inrixs-new-traffic-app-is-coming-for-you/** (accessed June 2018).

Greengard, S. "Delving into Gartner's 2016 Hype Cycle." *Baseline*, September 7, 2016.

Greig, J. "Gartner: AI Business Value Up 70% in 2018, and These Industries Will Benefit the Most." *Tech Republic*, April 25, 2018.

Haines, D. "Is Artificial Intelligence Making It Easier and Quicker to Get a New Job?" *Huffington Post UK*, December 4, 2017.

Hauari, G. "InsurersLeverage AI to Unlock Legacy Claims Data." *Information Management*, July 3, 2017.

Huang, G. "Why AI Doesn't Mean Taking the 'Human' Out of Human Resources." **Forbes.com**, September 27, 2017.

Hughes, T. "Google DeepMind's Program Beat Human at Go." *USA Today*, January 27, 2016.

ICAEW. "Artificial Intelligence and the Future of Accountancy." **artificial-intelligence-report.ashx/**, 2017.

Kaplan, J. *Artificial Intelligence: What Everyone Needs to Know*. London, UK: Oxford University Press, 2016.

Kharpal, A. "A.I. Is in a 'Golden Age' and Solving Problems That Were Once in the Realm of Sci-Fi, Jeff Bezos Says." *CNBC News*, May 8, 2017.

Kiron, D. "What Managers Need to Know About Artificial Intelligence?" *MITSloan Management Review*, January 25, 2017.

Knight, W. "Walmart's Robotic Shopping Carts Are the Latest Sign That Automation Is Eating Commerce." *Technology Review*, June 15, 2016.

Kolbjørnsrud, V., R. Amico, and R. J. Thomas. "How Artificial Intelligence Will Redefine Management." *Harvard Business Review*, November 2, 2016.

Korosec, K. "Inrix Updates Traffic App to Learn Your Daily Habits." *Fortune Tech*, March 30, 2016.

Liao, P-H., et al. "Applying Artificial Intelligence Technology to Support Decision-Making in Nursing: A Case Study in Taiwan." *Health Informatics Journal*, June 2015.

Marr, B., "The Key Definitions of Artificial Intelligence That Explain Its Importance." *Forbes*, February 14, 2018.

Marr, B. "What Everyone Should Know About Cognitive Computing." **Forbes.com**, March 23, 2016.

Martin, J. "10 Things Marketers Need to Know about AI." *CIO.com*, February 13, 2017.

McPherson, S.S. *Artificial Intelligence: Building Smarter Machines*. Breckenridge, CO: Twenty-First Century Books, 2017.

Meister, J. "The Future of Work: How Artificial Intelligence Will Transform the Employee Experience." **Forbes.com**, November 9, 2017.

Metz, C. "Facebook's Augmented Reality Engine Brings AI Right to Your Phone." *Wired*, April 19, 2017.

Mittal, V. "Top 15 Deep Learning Applications That Will Rule the World in 2018 and Beyond." **Medium.com**, October 3, 2017.

Narayan, K. "Leverage Artificial Intelligence to Build your Sales Pipeline." *LinkedIn*, February 14, 2018.

Ng, A. "What Artificial Intelligence Can and Can't Do Right Now." *Harvard Business Review*, November 9, 2016.

Nordrum, A. "Hedge Funds Look to Machine Learning, Crowdsourcing for Competitive Advantage." *IEEE Spectrum*, June 28, 2017.

Ovaska-Few, S. "How Artificial Intelligence Is Changing Accounting." *Journal of Accountancy*, October 9, 2017.

Padmanabhan, G. "Industry-Specific Augmented Intelligence: A Catalysts for AI in the Enterprise." *Forbes*, January 4, 2018.

Pennington, R. "Artificial Intelligence: The New Tool for Accomplishing an Old Goal in Marketing." *Huffington Post*, January 16, 2018.

Press, G. "Top 10 Hot Artificial Intelligence (AI) Technologies." *Forbes*, January 23, 2017.

Pyle, D., and C. San José. "An Executive's Guide to Machine Learning." McKinsey & Company, June 2015.

Reinharz, S. *An Introduction to Artificial Intelligence: Professional Edition: An Introductory Guide to the Evolution of Artificial Intelligence*. Kindle Edition. Seattle, WA: Simultaneous Device Usage (Amazon Digital Service), 2017.

Sample, I. "AI Watchdog Needed to Regulate Automated Decision-Making, Say Experts." *The Guardian*, January 27, 2017.

Santana, D. "Metromile Launches AI Claims Platform." *Digital Insurance*, July 25, 2017.

Savar, A. "3 Ways That A.I. Is Transforming HR and Recruiting." **INC.com**, June 26, 2017.

Schrage, M. "4 Models for Using AI to Make Decisions." *Harvard Business Review*, January 27, 2017.

Shah, J. "Robots Are Learning Complex Tasks Just by Watching Humans Do Them." *Harvard Business Review*, June 21, 2016.

Sharma, G. "China Unveils Multi-Billion Dollar Artificial Intelligence Plan." *International Business Times*, July 20, 2017. **ibtimes.co.uk/china-unveils-multi-billion-dollar-artificial-intelligence-plan-1631171/** (accessed January 2018).

Sincavage, D. "How Artificial Intelligence Will Change Decision-Making for Businesses." *Business 2 Community*, August 24, 2017.

Singh, H. "How Artificial Intelligence Will Transform Financial Services." *Information Management*, June 6, 2017.

SMBWorld Asia Editors. "Hays: Artificial Intelligence Set to Revolutionize Recruitment." *Enterprise Innovation*, August 30, 2017.

Smith, J. *Machine Learning: Machine Learning for Beginners. Can Machines Really Learn Like Humans? All About Artificial Intelligence (AI), Deep Learning and Digital Neural Networks*. Kindle Edition. Seattle, WA: Amazon Digital Service, 2017.

Staff. "Assisted, Augmented and Autonomous: The 3 Flavours of AI Decisions." *Software and Technology*, June 28, 2017. **tgdaily.com/technology/assisted-augmented-and-autonomous-the-3-flavours-of-ai-decisions**

Steffi, S. "List of 50 Unique AI Technologies." *Hacker* **Noon.com**, October 18, 2017.

Taylor, P. "Welcome to the Machine – Learning." *Forbes BrandVoice*, June 3, 2016. **forbes.com/sites/sap/2016/06/03/welcome-to-the-machine-learning/#3175d50940fe** (accessed June 2017).

Theobald, O. *Machine Learning for Absolute Beginners: A Plain English Introduction*. Kindle Edition. Seattle, WA, 2017.

Tiwan, R. "Artificial Intelligence (AI) in Banking Case Study Report 2017." *iCrowd Newswire*, July 7, 2017.

USC. "AI Computer Vision Breakthrough IDs Poachers in Less Than Half a Second." *Press Release*, February 8, 2018.

Violino, B. "Most Firms Expect Rapid Returns on Artificial Intelligence Investments." *Information Management*, November 1, 2017.

Warawa, J. "Here's Why Accountants (Yes, YOU!) Should Be Driving AI Innovation." *CPA Practice Advisor*, November 1, 2017.

Waxer, C. "Get Ready for the BOT Revolution." *Computerworld*, October 17, 2016.

Wellers, D., et al. "8 Ways Machine Learning Is Improving Companies' Work Processes." *Harvard Business Review*, May 31, 2017.

Wislow, E. "5 Ways to Use Artificial Intelligence (AI) in Human Resources." *Big Data Made Simple*, October 24, 2017. **bigdata-madesimple.com/5-ways-to-use-artificial-intelligence-ai-in-human-resources/**.

Yurcan, B. "TD's Innovation Agenda: Experiments with Alexa, AI and Augmented Reality." *Information Management*, December 27, 2017.

Zarkadakis, G. *In Our Own Image: Savior or Destroyer? The History and Future of Artificial Intelligence*. New York, NY: Pegasus Books, 2016.

Zhou, A. "EY, Deloitte and PwC Embrace Artificial Intelligence for Tax and Accounting." *Forbes.com*, November 14, 2017.

CHAPTER 3

Nature of Data, Statistical Modeling, and Visualization

LEARNING OBJECTIVES

- Understand the nature of data as they relate to business intelligence (BI) and analytics
- Learn the methods used to make real-world data analytics ready
- Describe statistical modeling and its relationship to business analytics
- Learn about descriptive and inferential statistics
- Define business reporting and understand its historical evolution
- Understand the importance of data/information visualization
- Learn different types of visualization techniques
- Appreciate the value that visual analytics brings to business analytics
- Know the capabilities and limitations of dashboards

In the age of Big Data and business analytics in which we are living, the importance of data is undeniable. Newly coined phrases such as "data are the oil," "data are the new bacon," "data are the new currency," and "data are the king" are further stressing the renewed importance of data. But the type of data we are talking about is obviously not just any data. The "garbage in garbage out—GIGO" concept/principle applies to today's Big Data phenomenon more so than any data definition that we have had in the past. To live up to their promise, value proposition, and ability to turn into insight, data have to be carefully created/identified, collected, integrated, cleaned, transformed, and properly contextualized for use in accurate and timely decision making.

Data are the main theme of this chapter. Accordingly, the chapter starts with a description of the nature of data: what they are, what different types and forms they can come in, and how they can be preprocessed and made ready for analytics. The first few sections of the chapter are dedicated to a deep yet necessary understanding and processing of data. The next few sections describe the statistical methods used to prepare data as input to produce both descriptive and inferential measures. Following the statistics sections are sections on reporting and visualization. A report is a communication artifact

prepared with the specific intention of converting data into information and knowledge and relaying that information in an easily understandable/digestible format. Today, these reports are visually oriented, often using colors and graphical icons that collectively look like a dashboard to enhance the information content. Therefore, the latter part of the chapter is dedicated to subsections that present the design, implementation, and best practices regarding information visualization, storytelling, and information dashboards.

This chapter has the following sections:

3.1 OPENING VIGNETTE: SiriusXM Attracts and Engages a New Generation of Radio Consumers with Data-Driven Marketing

SiriusXM Radio is a satellite radio powerhouse, the largest radio company in the world with $3.8 billion in annual revenues and a wide range of hugely popular music, sports, news, talk, and entertainment stations. The company, which began broadcasting in 2001 with 50,000 subscribers, had 18.8 million subscribers in 2009, and today has nearly 29 million.

Much of SiriusXM's growth to date is rooted in creative arrangements with automobile manufacturers; today, nearly 70 percent of new cars are SiriusXM enabled. Yet the company's reach extends far beyond car radios in the United States to a worldwide presence on the Internet, on smartphones, and through other services and distribution channels, including SONOS, JetBlue, and Dish.

BUSINESS CHALLENGE

Despite these remarkable successes, changes in customer demographics, technology, and a competitive landscape over the past few years have posed a new series of business challenges and opportunities for SiriusXM. Here are some notable ones:

- As its market penetration among new cars increased, the demographics of its buyers changed, skewing toward younger people with less discretionary income. How could SiriusXM reach this new demographic?
- As new cars become used cars and change hands, how could SiriusXM identify, engage, and convert second owners to paying customers?
- With its acquisition of the connected vehicle business from Agero—the leading provider of telematics in the U.S. car market—SiriusXM gained the ability to deliver its service via satellite and wireless networks. How could it successfully use this acquisition to capture new revenue streams?

PROPOSED SOLUTION: SHIFTING THE VISION TOWARD DATA-DRIVEN MARKETING

SiriusXM recognized that to address these challenges, it would need to become a high-performance, data-driven marketing organization. The company began making that shift by establishing three fundamental tenets. First, personalized interactions—not mass marketing—would rule the day. The company quickly understood that to conduct more personalized marketing, it would have to draw on past history and interactions as well as on a keen understanding of the consumer's place in the subscription life cycle.

Second, to gain that understanding, information technology (IT) and its external technology partners would need the ability to deliver integrated data, advanced analytics, integrated marketing platforms, and multichannel delivery systems.

And third, the company could not achieve its business goals without an integrated and consistent point of view across the company. Most important, the technology and business sides of SiriusXM would have to become true partners to best address the challenges involved in becoming a high-performance marketing organization that draws on data-driven insights to speak directly with consumers in strikingly relevant ways.

Those data-driven insights, for example, would enable the company to differentiate between consumers, owners, drivers, listeners, and account holders. The insights would help SiriusXM to understand what other vehicles and services are part of each household and create new opportunities for engagement. In addition, by constructing a coherent and reliable 360-degree view of all its consumers, SiriusXM could ensure that all messaging in all campaigns and interactions would be tailored, relevant, and consistent across all channels. The important bonus is that a more tailored and effective marketing is typically more cost-efficient.

IMPLEMENTATION: CREATING AND FOLLOWING THE PATH TO HIGH-PERFORMANCE MARKETING

At the time of its decision to become a high-performance marketing company, SiriusXM was working with a third-party marketing platform that did not have the capacity to support SiriusXM's ambitions. The company then made an important, forward-thinking decision to bring its marketing capabilities in-house—and then carefully plotted what it would need to do to make the transition successfully.

1. Improve data cleanliness through improved master data management and governance. Although the company was understandably impatient to put ideas into action, data hygiene was a necessary first step to create a reliable window into consumer behavior.
2. Bring marketing analytics in-house and expand the data warehouse to enable scale and fully support integrated marketing analytics.
3. Develop new segmentation and scoring models to run in databases, eliminating latency and data duplication.
4. Extend the integrated data warehouse to include marketing data and scoring, leveraging in-database analytics.
5. Adopt a marketing platform for campaign development.
6. Bring all of its capability together to deliver real-time offer management across all marketing channels: call center, mobile, Web, and in-app.

Completing those steps meant finding the right technology partner. SiriusXM chose Teradata because its strengths were a powerful match for the project and company. Teradata offered the ability to:

- Consolidate data sources with an integrated data warehouse (IDW), advanced analytics, and powerful marketing applications.
- Solve data-latency issues.

- Significantly reduce data movement across multiple databases and applications.
- Seamlessly interact with applications and modules for all marketing areas.
- Scale and perform at very high levels for running campaigns and analytics in-database.
- Conduct real-time communications with customers.
- Provide operational support, either via the cloud or on premise.

This partnership has enabled SiriusXM to move smoothly and swiftly along its road map, and the company is now in the midst of a transformational, five-year process. After establishing its strong data governance process, SiriusXM began by implementing its IDW, which allowed the company to quickly and reliably operationalize insights throughout the organization.

Next, the company implemented Customer Interaction Manager—part of the Teradata Integrated Marketing Cloud, which enables real-time, dialog-based customer interaction across the full spectrum of digital and traditional communication channels. SiriusXM also will incorporate the Teradata Digital Messaging Center.

Together, the suite of capabilities allows SiriusXM to handle direct communications across multiple channels. This evolution will enable real-time offers, marketing messages, and recommendations based on previous behavior.

In addition to streamlining the way it executes and optimizes outbound marketing activities, SiriusXM is also taking control of its internal marketing operations with the implementation of Marketing Resource Management, also part of the Teradata Integrated Marketing Cloud. The solution will allow SiriusXM to streamline workflow, optimize marketing resources, and drive efficiency through every penny of its marketing budget.

RESULTS: REAPING THE BENEFITS

As SiriusXM continues its evolution into a high-performance marketing organization, it already is benefiting from its thoughtfully executed strategy. Household-level consumer insights and a complete view of marketing touch strategy with each consumer enable SiriusXM to create more targeted offers at the household, consumer, and device levels. By bringing the data and marketing analytics capabilities in-house, SiriusXM achieved the following:

- Campaign results in near real time rather than four days, resulting in massive reductions in cycle times for campaigns and the analysts who support them.
- Closed-loop visibility, allowing the analysts to support multistage dialogs and in-campaign modifications to increase campaign effectiveness.
- Real-time modeling and scoring to increase marketing intelligence and sharpen campaign offers and responses at the speed of their business.

Finally, SiriusXM's experience has reinforced the idea that high-performance marketing is a constantly evolving concept. The company has implemented both processes and the technology that give it the capacity for continued and flexible growth.

QUESTIONS FOR THE OPENING VIGNETTE

1. What does SiriusXM do? In what type of market does it conduct its business?
2. What were its challenges? Comment on both technology and data-related challenges.
3. What were the proposed solutions?
4. How did the company implement the proposed solutions? Did it face any implementation challenges?
5. What were the results and benefits? Were they worth the effort/investment?
6. Can you think of other companies facing similar challenges that can potentially benefit from similar data-driven marketing solutions?

WHAT WE CAN LEARN FROM THIS VIGNETTE

Striving to thrive in a fast-changing competitive industry, SiriusXM realized the need for a new and improved marketing infrastructure (one that relies on data and analytics) to effectively communicate its value proposition to its existing and potential customers. As is the case in any industry, success or mere survival in entertainment depends on intelligently sensing the changing trends (likes and dislikes) and putting together the right messages and policies to win new customers while retaining the existing ones. The key is to create and manage successful marketing campaigns that resonate with the target population of customers and have a close feedback loop to adjust and modify the message to optimize the outcome. At the end, it was all about the precision in the way that SiriusXM conducted business: being proactive about the changing nature of the clientele and creating and transmitting the right products and services in a timely manner using a fact-based/data-driven holistic marketing strategy. Source identification, source creation, access and collection, integration, cleaning, transformation, storage, and processing of relevant data played a critical role in SiriusXM's success in designing and implementing a marketing analytics strategy as is the case in any analytically savvy successful company today, regardless of the industry in which they are participating.

Sources: C. Quinn, "Data-Driven Marketing at SiriusXM," Teradata Articles & News, 2016. **http://bigdata.teradata.com/US/Articles-News/Data-Driven-Marketing-At-SiriusXM/** (accessed August 2016); "SiriusXM Attracts and Engages a New Generation of Radio Consumers." **http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB8597.pdf?processed=1**.

3.2 NATURE OF DATA

Data are the main ingredient for any BI, data science, and business analytics initiative. In fact, they can be viewed as the raw material for what popular decision technologies produce—information, insight, and **knowledge**. Without data, none of these technologies could exist and be popularized—although traditionally we have built analytics models using expert knowledge and experience coupled with very little or no data at all; however, those were the old days, and now data are of the essence. Once perceived as a big challenge to collect, store, and manage, data today are widely considered among the most valuable assets of an organization with the potential to create invaluable insight to better understand customers, competitors, and the business processes.

Data can be small or very large. They can be structured (nicely organized for computers to process), or they can be unstructured (e.g., text that is created for humans and hence not readily understandable/consumable by computers). Data can come in small batches continuously or can pour in all at once as a large batch. These are some of the characteristics that define the inherent nature of today's data, which we often call Big Data. Even though these characteristics of data make them more challenging to process and consume, they also make the data more valuable because the characteristics enrich them beyond their conventional limits, allowing for the discovery of new and novel knowledge. Traditional ways to manually collect data (via either surveys or human-entered business transactions) mostly left their places to modern-day data collection mechanisms that use Internet and/or sensor/radio frequency identification (RFID)–based computerized networks. These automated data collection systems are not only enabling us to collect more volumes of data but also enhancing the **data quality** and integrity. Figure 3.1 illustrates a typical analytics continuum—data to analytics to actionable information.

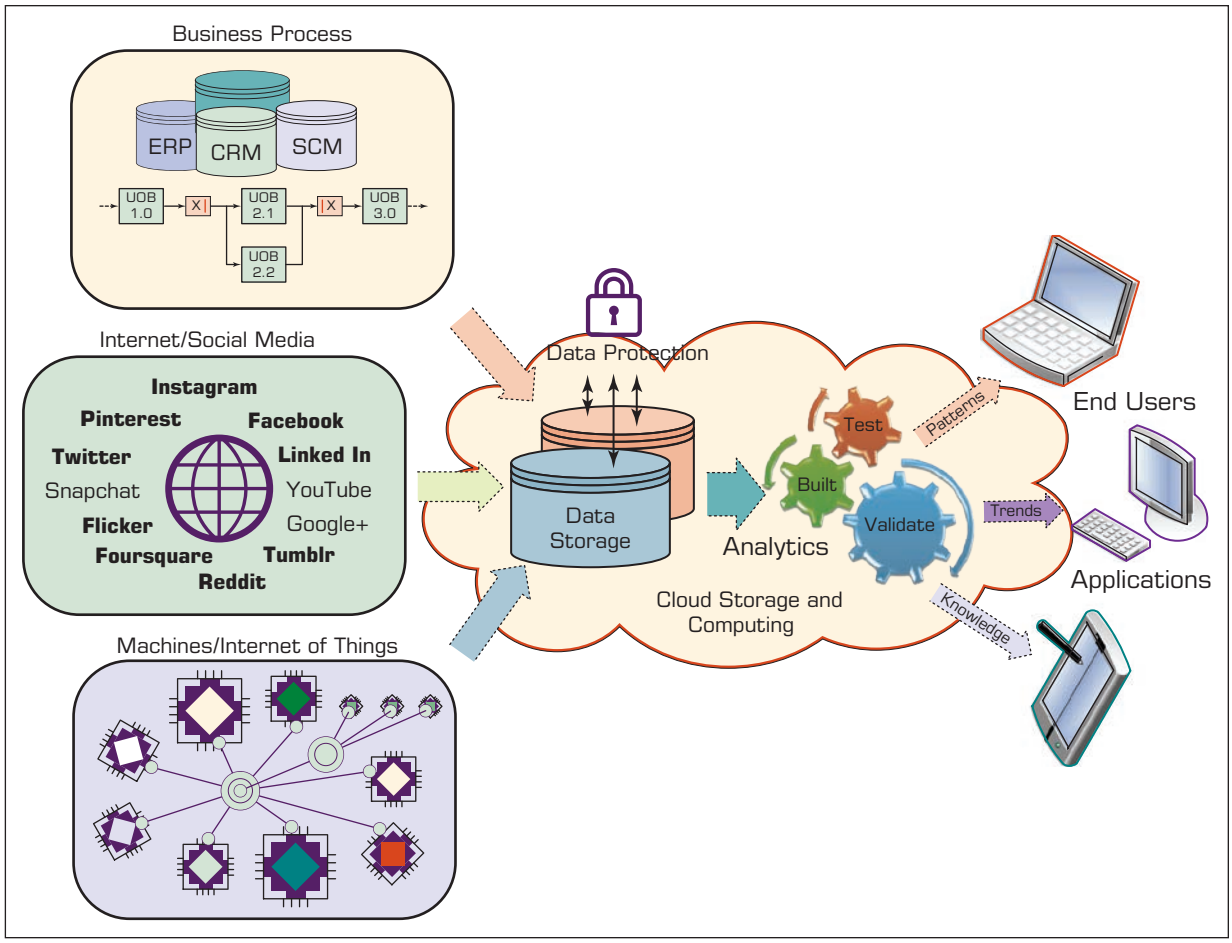


FIGURE 3.1 **A Data to Knowledge Continuum.**

Although their value proposition is undeniable, to live up their promise, data must comply with some basic usability and quality metrics. Not all data are useful for all tasks, obviously. That is, data must match with (have the coverage of the specifics for) the task for which they are intended to be used. Even for a specific task, the relevant data on hand need to comply with the quality and quantity requirements. Essentially, data have to be analytics ready. So what does it mean to make data analytics ready? In addition to its relevancy to the problem at hand and the quality/quantity requirements, it also has to have a certain structure in place with key fields/variables with properly normalized values. Furthermore, there must be an organization-wide agreed-on definition for common variables and subject matters (sometimes also called *master data management*), such as how to define a customer (what characteristics of customers are used to produce a holistic enough representation to analytics) and where in the business process the customer-related information is captured, validated, stored, and updated.

Sometimes the representation of the data depends on the type of analytics being employed. Predictive algorithms generally require a flat file with a target variable, so making data **analytics ready** for prediction means that data sets must be transformed into a flat-file format and made ready for ingestion into those predictive algorithms. It is also imperative to match the data to the needs and wants of a specific predictive algorithm and/or a software tool. For instance, neural network algorithms require all input variables

to be numerically represented (even the nominal variables need to be converted into pseudo binary numeric variables), whereas decision tree algorithms do not require such numerical transformation—they can easily and natively handle a mix of nominal and numeric variables.

Analytics projects that overlook data-related tasks (some of the most critical steps) often end up with the wrong answer for the right problem, and these unintentionally created, seemingly good answers could lead to inaccurate and untimely decisions. Following are some of the characteristics (metrics) that define the readiness level of data for an analytics study (Delen, 2015; Kock, McQueen, & Corner, 1997).

- **Data source reliability.** This term refers to the originality and appropriateness of the storage medium where the data are obtained—answering the question of "Do we have the right confidence and belief in this data source?" If at all possible, one should always look for the original source/creator of the data to eliminate/mitigate the possibilities of data misrepresentation and data transformation caused by the mishandling of the data as they moved from the source to destination through one or more steps and stops along the way. Every move of the data creates a chance to unintentionally drop or reformat data items, which limits the integrity and perhaps true accuracy of the data set.
- **Data content accuracy.** This means that data are correct and are a good match for the analytics problem—answering the question of "Do we have the right data for the job?" The data should represent what was intended or defined by the original source of the data. For example, the customer's contact information recorded within a database should be the same as what the customer said it was. Data accuracy will be covered in more detail in the following subsection.
- **Data accessibility.** This term means that the data are easily and readily obtainable—answering the question of "Can we easily get to the data when we need to?" Access to data can be tricky, especially if they are stored in more than one location and storage medium and need to be merged/transformed while accessing and obtaining them. As the traditional relational database management systems leave their place (or coexist with a new generation of data storage mediums such as data lakes and Hadoop infrastructure), the importance/criticality of data accessibility is also increasing.
- **Data security and data privacy. Data security** means that the data are secured to allow only those people who have the authority and the need to access them and to prevent anyone else from reaching them. Increasing popularity in educational degrees and certificate programs for Information Assurance is evidence of the criticality and the increasing urgency of this data quality metric. Any organization that maintains health records for individual patients must have systems in place that not only safeguard the data from unauthorized access (which is mandated by federal laws such as the Health Insurance Portability and Accountability Act [HIPAA]) but also accurately identify each patient to allow proper and timely access to records by authorized users (Annas, 2003).
- **Data richness.** This means that all required data elements are included in the data set. In essence, richness (or comprehensiveness) means that the available variables portray a rich enough dimensionality of the underlying subject matter for an accurate and worthy analytics study. It also means that the information content is complete (or near complete) to build a predictive and/or prescriptive analytics model.
- **Data consistency.** This means that the data are accurately collected and combined/merged. Consistent data represent the dimensional information (variables of interest) coming from potentially disparate sources but pertaining to the same subject. If the data integration/merging is not done properly, some of the variables of different subjects could appear in the same record—having two different patient

records mixed up; for instance, this could happen while merging the demographic and clinical test result data records.

- **Data currency/data timeliness.** This means that the data should be up-to-date (or as recent/new as they need to be) for a given analytics model. It also means that the data are recorded at or near the time of the event or observation so that the time delay–related misrepresentation (incorrectly remembering and encoding) of the data is prevented. Because accurate analytics relies on accurate and timely data, an essential characteristic of analytics-ready data is the timeliness of the creation and access to data elements.
- **Data granularity.** This requires that the variables and data values be defined at the lowest (or as low as required) level of detail for the intended use of the data. If the data are aggregated, they might not contain the level of detail needed for an analytics algorithm to learn how to discern different records/cases from one another. For example, in a medical setting, numerical values for laboratory results should be recorded to the appropriate decimal place as required for the meaningful interpretation of test results and proper use of those values within an analytics algorithm. Similarly, in the collection of demographic data, data elements should be defined at a granular level to determine the differences in outcomes of care among various subpopulations. One thing to remember is that the data that are aggregated cannot be disaggregated (without access to the original source), but they can easily be aggregated from its granular representation.
- **Data validity.** This is the term used to describe a match/mismatch between the actual and expected data values of a given variable. As part of data definition, the acceptable values or value ranges for each data element must be defined. For example, a valid data definition related to gender would include three values: male, female, and unknown.
- **Data relevancy.** This means that the variables in the data set are all relevant to the study being conducted. Relevancy is not a dichotomous measure (whether a variable is relevant or not); rather, it has a spectrum of relevancy from least relevant to most relevant. Based on the analytics algorithms being used, one can choose to include only the most relevant information (i.e., variables) or, if the algorithm is capable enough to sort them out, can choose to include all the relevant ones regardless of their levels. One thing that analytics studies should avoid is including totally irrelevant data into the model building because this could contaminate the information for the algorithm, resulting in inaccurate and misleading results.

The above-listed characteristics are perhaps the most prevailing metrics to keep up with; the true data quality and excellent analytics readiness for a specific application domain would require different levels of emphasis to be placed on these metric dimensions and perhaps add more specific ones to this collection. The following section will delve into the nature of data from a taxonomical perspective to list and define different data types as they relate to different analytics projects.

SECTION 3.2 REVIEW QUESTIONS

1. How do you describe the importance of data in analytics? Can we think of analytics without data?
2. Considering the new and broad definition of business analytics, what are the main inputs and outputs to the analytics continuum?
3. Where do the data for business analytics come from?
4. In your opinion, what are the top three data-related challenges for better analytics?
5. What are the most common metrics that make for analytics-ready data?

3.3 SIMPLE TAXONOMY OF DATA

The term *data* (**datum** in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. Data can consist of numbers, letters, words, images, voice recordings, and so on, as measurements of a set of variables (characteristics of the subject or event that we are interested in studying). Data are often viewed as the lowest level of abstraction from which information and then knowledge is derived.

At the highest level of abstraction, one can classify data as structured and unstructured (or semistructured). **Unstructured data**/semistructured data are composed of any combination of textual, imagery, voice, and Web content. Unstructured/semistructured data will be covered in more detail in the text mining and Web mining chapter. **Structured data** are what data mining algorithms use and can be classified as categorical or numeric. The **categorical data** can be subdivided into nominal or **ordinal data**, whereas numeric data can be subdivided into intervals or ratios. Figure 3.2 shows a simple **data taxonomy**.

- **Categorical data.** These represent the labels of multiple classes used to divide a variable into specific groups. Examples of categorical variables include race, sex, age group, and educational level. Although the latter two variables can also be considered in a numerical manner by using exact values for age and highest grade completed, for example, it is often more informative to categorize such variables into a relatively small number of ordered classes. The categorical data can also be called *discrete data*, implying that they represent a finite number of values with no continuum between them. Even if the values used for the categorical (or discrete) variables are numeric, these numbers are nothing more than symbols and do not imply the possibility of calculating fractional values.
- **Nominal data.** These contain measurements of simple codes assigned to objects as labels, which are not measurements. For example, the variable *marital status* can be generally categorized as (1) single, (2) married, and (3) divorced. **Nominal**

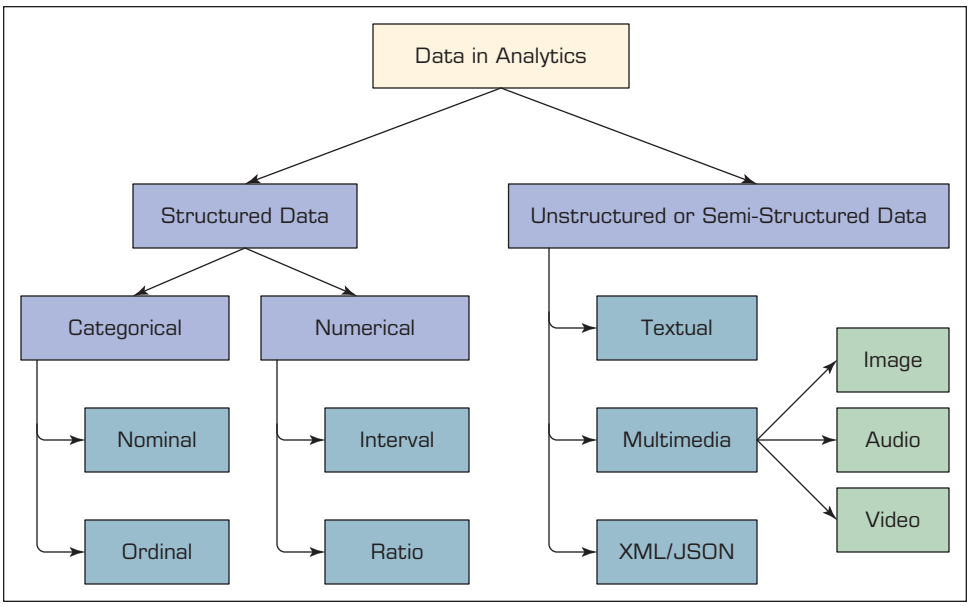


FIGURE 3.2 **A Simple Taxonomy of Data.**

data can be represented with binomial values having two possible values (e.g., yes/no, true/false, good/bad) or multinomial values having three or more possible values (e.g., brown/green/blue, white/black/Latino/Asian, single/married/divorced).

- **Ordinal data.** These contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable *credit score* can be generally categorized as (1) low, (2) medium, or (3) high. Similar ordered relationships can be seen in variables such as age group (i.e., child, young, middle-aged, elderly) and educational level (i.e., high school, college, graduate school). Some predictive analytic algorithms, such as *ordinal multiple logistic regression*, take into account this additional rank-order information to build a better classification model.
- **Numeric data.** These represent the numeric values of specific variables. Examples of numerically valued variables include age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees). Numeric values representing a variable can be integers (only whole numbers) or real (also fractional numbers). The numeric data can also be called *continuous data*, implying that the variable contains continuous measures on a specific scale that allows insertion of interim values. Unlike a discrete variable, which represents finite, countable data, a continuous variable represents scalable measurements, and it is possible for the data to contain an infinite number of fractional values.
- **Interval data.** These are variables that can be measured on interval scales. A common example of interval scale measurement is temperature on the Celsius scale. In this particular scale, the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water in atmospheric pressure; that is, there is not an absolute zero value.
- **Ratio data.** These include measurement variables commonly found in the physical sciences and engineering. Mass, length, time, plane angle, energy, and electric charge are examples of physical measures that are ratio scales. The scale type takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. Informally, the distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value. For example, the Kelvin temperature scale has a nonarbitrary zero point of absolute zero, which is equal to –273.15 degrees Celsius. This zero point is nonarbitrary because the particles that comprise matter at this temperature have zero kinetic energy.

Other data types, including textual, spatial, imagery, video, and voice, need to be converted into some form of categorical or numeric representation before they can be processed by analytics methods (data mining algorithms; Delen, 2015). Data can also be classified as static or dynamic (i.e., temporal or time series).

Some predictive analytics (i.e., data mining) methods and machine-learning algorithms are very selective about the type of data that they can handle. Providing them with incompatible data types can lead to incorrect models or (more often) halt the model development process. For example, some data mining methods need all the variables (both input and output) represented as numerically valued variables (e.g., neural networks, support vector machines, logistic regression). The nominal or ordinal variables are converted into numeric representations using some type of *1-of-N* pseudo variables (e.g., a categorical variable with three unique values can be transformed into three pseudo variables with binary values—1 or 0). Because this process could increase the number of variables, one should be cautious about the effect of such representations, especially for the categorical variables that have large numbers of unique values.

Similarly, some predictive analytics methods, such as ID3 (a classic decision tree algorithm) and rough sets (a relatively new rule induction algorithm), need all the variables represented as categorically valued variables. Early versions of these methods required the user to discretize numeric variables into categorical representations before they could be processed by the algorithm. The good news is that most implementations of these algorithms in widely available software tools accept a mix of numeric and nominal variables and internally make the necessary conversions before processing the data.

Data come in many different variable types and representation schemas. Business analytics tools are continuously improving in their ability to help data scientists in the daunting task of data transformation and data representation so that the data requirements of specific predictive models and algorithms can be properly executed. Application Case 3.1 illustrates a business scenario in which one of the largest telecommunication companies streamlined and used a wide variety of rich data sources to generate customers insight to prevent churn and to create new revenue sources.

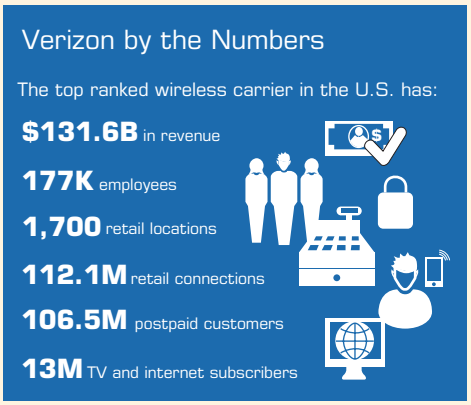
Application Case 3.1

Verizon Answers the Call for Innovation: The Nation's Largest Network Provider uses Advanced Analytics to Bring the Future to its Customers

The Problem

In the ultra-competitive telecommunications industry, staying relevant to consumers while finding new sources of revenue is critical, especially since current revenue sources are in decline.

For Fortune 13 powerhouse Verizon, the secret weapon that catapulted the company into the nation's largest and most reliable network provider is also guiding the business toward future success (see the following figure for some numbers about Verizon). The secret weapon? Data and analytics. Because telecommunication companies are typically rich in data, having the right analytics solution and personnel in place can uncover critical insights that benefit every area of the business.



The Backbone of the Company

Since its inception in 2000, Verizon has partnered with Teradata to create a data and analytics architecture that drives innovation and science-based decision making. The goal is to stay relevant to customers while also identifying new business opportunities and making adjustments that result in more cost-effective operations.

"With business intelligence, we help the business identify new business opportunities or make course corrections to operate the business in a more cost-effective way," said Grace Hwang, executive director of Financial Performance & Analytics, BI, for Verizon. "We support decision makers with the most relevant information to improve the competitive advantage of Verizon."

By leveraging data and analytics, Verizon is able to offer a reliable network, ensure customer satisfaction, and develop products and services that consumers want to buy.

"Our incubator of new products and services will help bring the future to our customers," Hwang said. "We're using our network to make breakthroughs in

(*Continued*)

Application Case 3.1 (Continued)

interactive entertainment, digital media, the Internet of Things, and broadband services."

Data Insights across Three Business Units

Verizon relies on advanced analytics that are executed on the Teradata® Unified Data Architecture™ to support its business units. The analytics enable Verizon to deliver on its promise to help customers innovate their lifestyles and provide key insights to support these three areas:

- Identify new revenue sources. Research and development teams use data, analytics, and strategic partnerships to test and develop with the Internet of Things (IoT). The new frontier in data is IoT, which will lead to new revenues that in turn generate opportunities for top-line growth. Smart cars, smart agriculture, and smart IoT will all be part of this new growth.
- Predict churn in the core mobile business. Verizon has multiple use cases that demonstrate how its advanced analytics enable laser-accurate churn prediction—within a one to two percent margin—in the mobile space. For a $131 billion company, predicting churn with such precision is significant. By recognizing specific patterns in tablet data usage, Verizon can identify which customers most often access their tablets, then engage those who do not.
- Forecast mobile phone plans. Customer behavioral analytics allow finance to better predict earnings in fast-changing market conditions. The U.S. wireless industry is moving from monthly payments for both the phone and the service to paying for the phone independently. This opens up a new opportunity for Verizon to gain business. The analytic environment helps Verizon better predict churn with new plans and forecast the impact of changes to pricing plans.

The analytics deliver what Verizon refers to as "honest data" that inform various business units. "Our mission is to be the honest voice and the independent third-party opinion on the success or opportunities for improvement to the business," Hwang explains. "So my unit is viewed as the golden source of information, and we come across with the honest voice, and a lot of the business decisions are through various rungs of course correction."

Hwang adds that oftentimes, what forces a company to react is competitors affecting change in the marketplace, rather than the company making the wrong decisions. "So we try to guide the business through the best course of correction, wherever applicable, timely, so that we can continue to deliver record-breaking results year after year," she said. "I have no doubt that the business intelligence had led to such success in the past."

Disrupt and Innovate

Verizon leverages advanced analytics to optimize marketing by sending the most relevant offers to customers. At the same time, the company relies on analytics to ensure they have the financial acumen to stay number one in the U.S. mobile market. By continuing to disrupt the industry with innovative products and solutions, Verizon is positioned to remain the wireless standard for the industry.

"We need the marketing vision and the sales rigor to produce the most relevant offer to our customers, and then at the same time we need to have the finance rigor to ensure that whatever we offer to the customer is also profitable to the business so that we're responsible to our shareholders," Hwang says.

In Summary—Executing the Seven Ps of Modern Marketing

Telecommunications giant Verizon uses seven Ps to drive its modern-day marketing efforts. The Ps, when used in unison, help Verizon penetrate the market in the way it predicted.

1. **People**: Understanding customers and their needs to create the product.
2. **Place**: Where customers shop.
3. **Product**: The item that's been manufactured and is for sale.

4. **Process**: How customers get to the shop or place to buy the product.
5. **Pricing**: Working with promotions to get customers' attention.
6. **Promo**: Working with pricing to get customers' attention.
7. **Physical evidence**: The business intelligence that gives insights.

"The Aster and Hadoop environment allows us to explore things we suspect could be the reasons for breakdown in the seven Ps," says Grace Hwang, executive director of Financial Performance & Analytics, BI, for Verizon. "This goes back to providing the business value to our decision-makers. With each step in the seven Ps, we ought to be able to tell them where there are opportunities for improvement."

Questions for Case 3.1

1. What was the challenge Verizon was facing?
2. What was the data-driven solution proposed for Verizon's business units?
3. What were the results?

Source: Teradata Case Study "Verizon Answers the Call for Innovation" **https://www.teradata.com/Resources/Case-Studies/Verizon-answers-the-call-for-innovation** (accessed July 2018).

SECTION 3.3 REVIEW QUESTIONS

1. What are data? How do data differ from information and knowledge?
2. What are the main categories of data? What types of data can we use for BI and analytics?
3. Can we use the same data representation for all analytics models? Why, or why not?
4. What is a *1-of-N* data representation? Why and where is it used in analytics?

3.4 ART AND SCIENCE OF DATA PREPROCESSING

Data in their original form (i.e., the real-world data) are not usually ready to be used in analytics tasks. They are often dirty, misaligned, overly complex, and inaccurate. A tedious and time-demanding process (so-called **data preprocessing**) is necessary to convert the raw real-world data into a well-refined form for analytics algorithms (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Many analytics professionals would testify that the time spent on data preprocessing (which is perhaps the least enjoyable phase in the whole process) is significantly longer than the time spent on the rest of the analytics tasks (the fun of analytics model building and assessment). Figure 3.3 shows the main steps in the data preprocessing endeavor.

In the first step of data preprocessing, the relevant data are collected from the identified sources, the necessary records and variables are selected (based on an intimate understanding of the data, the unnecessary information is filtered out), and the records coming from multiple data sources are integrated/merged (again, using the intimate understanding of the data, the synonyms and homonyms are able to be handled properly).

In the second step of data preprocessing, the data are cleaned (this step is also known as *data scrubbing*). Data in their original/raw/real-world form are usually dirty (Hernández & Stolfo, 1998; Kim et al., 2003). In this phase, the values in the data set are identified and dealt with. In some cases, missing values are an anomaly in the data set, in which case they need to be imputed (filled with a most probable value) or ignored; in other cases, the missing values are a natural part of the data set

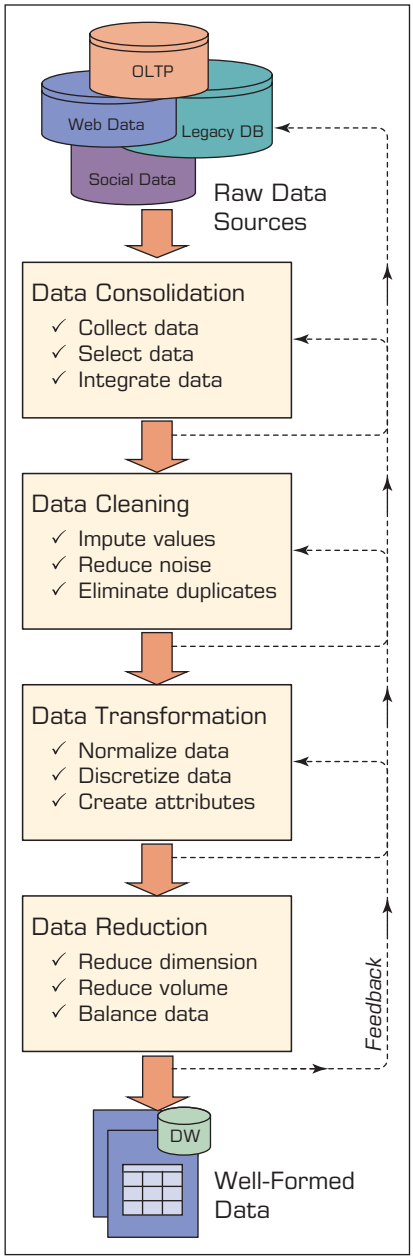


FIGURE 3.3 **Data Preprocessing Steps.**

(e.g., the *household income* field is often left unanswered by people who are in the top income tier). In this step, the analyst should also identify noisy values in the data (i.e., the outliers) and smooth them out. In addition, inconsistencies (unusual values within a variable) in the data should be handled using domain knowledge and/or expert opinion.

In the third step of data preprocessing, the data are transformed for better processing. For instance, in many cases, the data are normalized between a certain minimum and maximum for all variables to mitigate the potential bias of one variable having

large numeric values (such as household income) dominating other variables (such as *number of dependents* or *years in service,* which could be more important) having smaller values. Another transformation that takes place is discretization and/or aggregation. In some cases, the numeric variables are converted to categorical values (e.g., low, medium, high); in other cases, a nominal variable's unique value range is reduced to a smaller set using concept hierarchies (e.g., as opposed to using the individual states with 50 different values, one could choose to use several regions for a variable that shows location) to have a data set that is more amenable to computer processing. Still, in other cases, one might choose to create new variables based on the existing ones to magnify the information found in a collection of variables in the data set. For instance, in an organ transplantation data set, one might choose to use a single variable showing the blood-type match (1: match, 0: no match) as opposed to separate multinominal values for the blood type of both the donor and the recipient. Such simplification could increase the information content while reducing the complexity of the relationships in the data.

The final phase of data preprocessing is data reduction. Even though data scientists (i.e., analytics professionals) like to have large data sets, too much data can also be a problem. In the simplest sense, one can visualize the data commonly used in predictive analytics projects as a flat file consisting of two dimensions: variables (the number of columns) and cases/records (the number of rows). In some cases (e.g., image processing and genome projects with complex microarray data), the number of variables can be rather large, and the analyst must reduce the number to a manageable size. Because the variables are treated as different dimensions that describe the phenomenon from different perspectives, in predictive analytics and data mining, this process is commonly called **dimensional reduction** (or **variable selection**). Even though there is not a single best way to accomplish this task, one can use the findings from previously published literature; consult domain experts; run appropriate statistical tests (e.g., principal component analysis or independent component analysis); and, more preferably, use a combination of these techniques to successfully reduce the dimensions in the data into a more manageable and most relevant subset.

With respect to the other dimension (i.e., the number of cases), some data sets can include millions or billions of records. Even though computing power is increasing exponentially, processing such a large number of records cannot be practical or feasible. In such cases, one might need to sample a subset of the data for analysis. The underlying assumption of sampling is that the subset of the data will contain all relevant patterns of the complete data set. In a homogeneous data set, such an assumption could hold well, but real-world data are hardly ever homogeneous. The analyst should be extremely careful in selecting a subset of the data that reflects the essence of the complete data set and is not specific to a subgroup or subcategory. The data are usually sorted on some variable, and taking a section of the data from the top or bottom could lead to a biased data set on specific values of the indexed variable; therefore, always try to randomly select the records on the sample set. For skewed data, straightforward random sampling might not be sufficient, and stratified sampling (a proportional representation of different subgroups in the data is represented in the sample data set) might be required. Speaking of skewed data, it is a good practice to balance the highly skewed data by either oversampling the less represented or undersampling the more represented classes. Research has shown that balanced data sets tend to produce better prediction models than unbalanced ones (Thammasiri et al., 2014).

The essence of data preprocessing is summarized in Table 3.1, which maps the main phases (along with their problem descriptions) to a representative list of tasks and algorithms.

TABLE 3.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/max, mode, etc.); recode the missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalize the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or -1 to $+1$) by using a variety of normalization or scaling techniques.
	Discretize or aggregate the data	If needed, convert the numeric variables into discrete representations using range- or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Use principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Perform random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

It is almost impossible to underestimate the value proposition of data preprocessing. It is one of those time-demanding activities in which investment of time and effort pays off without a perceivable limit for diminishing returns. That is, the more resources you invest in it, the more you will gain at the end. Application Case 3.2 illustrates an interesting study that used raw, readily available academic data within an educational organization to develop predictive models to better understand attrition and improve freshman student retention in a large higher education institution. As the application case clearly states, each and every data preprocessing task described in Table 3.1 was critical to a successful execution of the underlying analytics project, especially the task that related to the balancing of the data set.

Application Case 3.2 Improving Student Retention with Data-Driven Analytics

Student attrition has become one of the most challenging problems for decision makers in academic institutions. Despite all the programs and services that are put in place to help retain students, according to the U.S. Department of Education's Center for Educational Statistics (**nces.ed.gov**), only about half of those who enter higher education actually earn a bachelor's degree. Enrollment management and the retention of students have become a top priority for administrators of colleges and universities in the United States and other countries around the world. High dropout of students usually results in overall financial loss, lower graduation rates, and an inferior school reputation in the eyes of all stakeholders. The legislators and policy makers who oversee higher education and allocate funds, the parents who pay for their children's education to prepare them for a better future, and the students who make college choices look for evidence of institutional quality and reputation to guide their decision-making processes.

The Proposed Solution

To improve student retention, one should try to understand the nontrivial reasons behind the attrition. To be successful, one should also be able to accurately identify those students who are at risk of dropping out. So far, the vast majority of student attrition research has been devoted to understanding this complex, yet crucial, social phenomenon. Even though these qualitative, behavioral, and survey-based studies revealed invaluable insight by developing and testing a wide range of theories, they do not provide the much-needed instruments to accurately predict (and potentially improve) student attrition. The project summarized in this case study proposed a quantitative research approach in which the historical institutional data from student databases could be used to develop models that are capable of predicting as well as explaining the institution-specific nature of the attrition problem. The proposed analytics approach is shown in Figure 3.4.

Although the concept is relatively new to higher education, for more than a decade now, similar problems in the field of marketing management have been studied using predictive data analytics techniques under the name of "churn analysis" where the purpose has been to identify a sample among current customers to answer the question, "Who among our current customers are more likely to stop buying our products or services?" so that some kind of mediation or intervention process can be executed to retain them. Retaining existing customers is crucial because, as we all know and as the related research has shown time and time again, acquiring a new customer costs on an order of magnitude more effort, time, and money than trying to keep the one that you already have.

Data Are of the Essence

The data for this research project came from a single institution (a comprehensive public university located in the Midwest region of the United States) with an average enrollment of 23,000 students, of which roughly 80 percent are the residents of the same state and roughly 19 percent of the students are listed under some minority classification. There is no significant difference between the two genders in the enrollment numbers. The average freshman student retention rate for the institution was about 80 percent, and the average six-year graduation rate was about 60 percent.

The study used five years of institutional data, which entailed 16,000+ students enrolled as freshmen, consolidated from various and diverse university student databases. The data contained variables related to students' academic, financial, and demographic characteristics. After merging and converting the multidimensional student data into a single flat file (a file with columns representing the variables and rows representing the student records), the resultant file was assessed and preprocessed to identify and remedy anomalies and unusable values. As an example, the study removed all international student records from the data set because they did not contain information about some of the most reputed predictors (e.g., high school GPA, SAT scores). In the data transformation phase, some of the variables were aggregated (e.g., "Major" and "Concentration" variables aggregated to binary variables MajorDeclared and ConcentrationSpecified)

(*Continued*)

Application Case 3.2 (Continued)

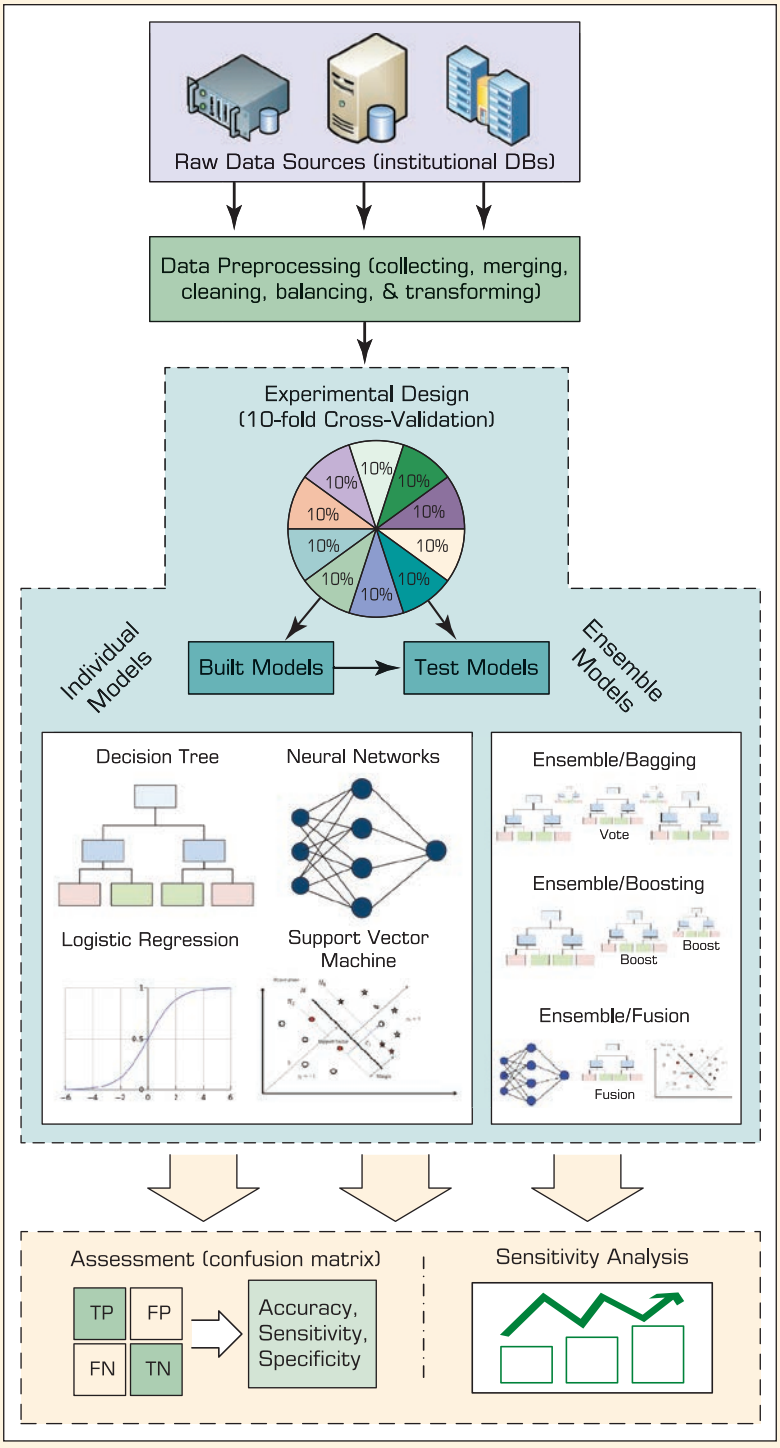


FIGURE 3.4 An Analytics Approach to Predicting Student Attrition.

for better interpretation for the predictive modeling. In addition, some of the variables were used to derive new variables (e.g., Earned/Registered ratio and YearsAfterHighSchool).

$$\text{Earned/Registered} = \text{EarnedHours}/\text{RegisteredHours}$$

$$\text{YearsAfterHighSchool} = \text{FreshmenEnrollmentYear} - \text{HighSchoolGraduationYear}$$

The *Earned/Registered* ratio was created to have a better representation of the students' resiliency and determination in their first semester of the freshman year. Intuitively, one would expect greater values for this variable to have a positive impact on retention/persistence. The *YearsAfterHighSchool* was created to measure the impact of the time taken between high school graduation and initial college enrollment. Intuitively, one would expect this variable to be a contributor to the prediction of attrition. These aggregations and derived variables are determined based on a number of experiments conducted for a number of logical hypotheses. The ones that made more common sense and the ones that led to better prediction accuracy were kept in the final variable set. Reflecting the true nature of the subpopulation (i.e., the freshmen students), the dependent variable (i.e., "Second Fall Registered") contained many more *yes* records (~80%) than *no* records (~20%; see Figure 3.5).

Research shows that having such imbalanced data has a negative impact on model performance. Therefore, the study experimented with the options of using and comparing the results of the same type of models built with the original imbalanced data (biased for the *yes* records) and the well-balanced data.

Modeling and Assessment

The study employed four popular classification methods (i.e., artificial neural networks, decision trees, support vector machines, and logistic regression) along with three model ensemble techniques (i.e., bagging, busting, and information fusion). The results obtained from all model types were then compared to each other using regular classification model assessment methods (e.g., overall predictive accuracy, sensitivity, specificity) on the holdout samples.

In machine-learning algorithms (some of which will be covered in Chapter 4), sensitivity analysis is a method for identifying the "cause-and-effect" relationship between the inputs and outputs of a given prediction model. The fundamental idea behind sensitivity analysis is that it measures the importance of predictor variables based on the change in modeling performance that occurs if a predictor variable is not included in the model. This modeling and experimentation practice is also called a leave-one-out assessment. Hence, the measure of sensitivity of a specific predictor variable is the ratio of the error of the trained model without the predictor variable to the error of the model that includes this predictor variable. The more sensitive

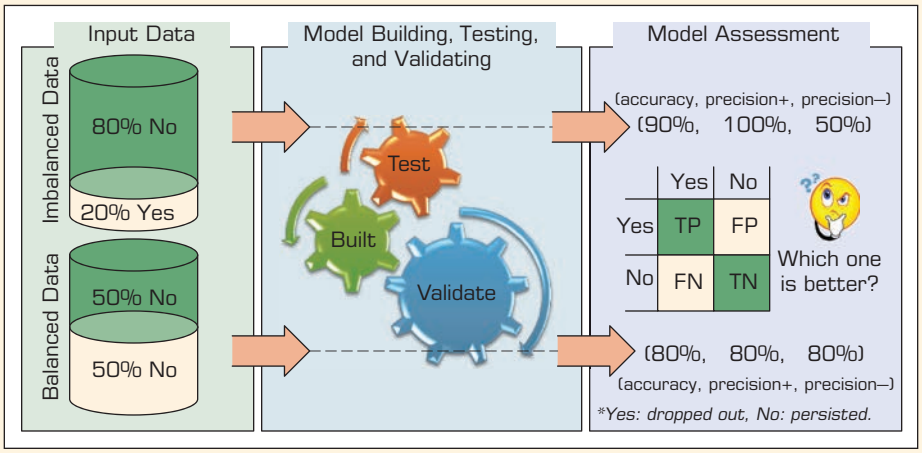


FIGURE 3.5 **A Graphical Depiction of the Class Imbalance Problem.**

(*Continued*)

Application Case 3.2 (Continued)

the network is to a particular variable, the greater the performance decrease would be in the absence of that variable and therefore the greater the ratio of importance. In addition to the predictive power of the models, the study also conducted sensitivity analyses to determine the relative importance of the input variables.

The Results

In the first set of experiments, the study used the original imbalanced data set. Based on the 10-fold cross-validation assessment results, the support vector machines produced the best accuracy with an overall prediction rate of 87.23 percent, and the decision tree was the runner-up with an overall prediction rate of 87.16 percent, followed by artificial neural networks and logistic regression with overall prediction rates of 86.45 percent and 86.12 percent, respectively (see Table 3.2). A careful examination of these results reveals that the prediction accuracy for the "Yes" class is significantly higher than the prediction accuracy of the "No" class. In fact, all four model types predicted the students who are likely to return for the second year with better than 90 percent accuracy, but the types did poorly on predicting the students who are likely to drop out after the freshman year with less than 50 percent accuracy. Because the prediction of the "No" class is the main purpose of this study, less than 50 percent accuracy for this class was deemed not acceptable. Such a difference in prediction accuracy of the two classes can (and should) be attributed to the imbalanced nature of the training data set (i.e., ~80% "Yes" and ~20% "No" samples).

The next round of experiments used a well-balanced data set in which the two classes are represented nearly equally in counts. In realizing this approach, the study took all samples from the minority class (i.e., the "No" class herein), randomly selected an equal number of samples from the majority class (i.e., the "Yes" class herein), and repeated this process 10 times to reduce potential bias of random sampling. Each of these sampling processes resulted in a data set of 7,000+ records, of which both class labels ("Yes" and "No") were equally represented. Again, using a 10-fold cross-validation methodology, the study developed and tested prediction models for all four model types. The results of these experiments are shown in Table 3.3. Based on the hold-out sample results, support vector machines once again generated the best overall prediction accuracy with 81.18 percent followed by decision trees, artificial neural networks, and logistic regression with an overall prediction accuracy of 80.65 percent, 79.85 percent, and 74.26 percent, respectively. As can be seen in the per-class accuracy figures, the prediction models did significantly better on predicting the "No" class with the well-balanced data than they did with the unbalanced data. Overall, the three machine-learning techniques performed significantly better than their statistical counterpart, logistic regression.

Next, another set of experiments was conducted to assess the predictive ability of the three ensemble models. Based on the 10-fold cross-validation methodology, the information fusion–type ensemble model produced the best results with an overall prediction rate of 82.10 percent, followed by the bagging-type ensembles and boosting-type

TABLE 3.2 Prediction Results for the Original/Unbalanced Data Set

	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1,494	384	1,518	304	1,478	255	1,438	376
Yes	1,596	11,142	1,572	11,222	1,612	11,271	1,652	11,150
SUM	3,090	11,526	3,090	11,526	3,090	11,526	3,090	11,526
Per-class accuracy	48.35%	96.67%	49.13%	97.36%	47.83%	97.79%	46.54%	96.74%
Overall accuracy	86.45%		87.16%		87.23%		86.12%	

*ANN: Artificial Neural Network; MLP: Multi-Layer Perceptron; DT: Decision Tree; SVM: Support Vector Machine; LR: Logistic Regression

TABLE 3.3 Prediction Results for the Balanced Data Set

Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2,309	464	2311	417	2,313	386	2,125	626
Yes	781	2,626	779	2,673	777	2,704	965	2,464
SUM	3,090	3,090	3,090	3,090	3,090	3,090	3,090	3,090
Per-class accuracy	74.72%	84.98%	74.79%	86.50%	74.85%	87.51%	68.77%	79.74%
Overall accuracy	79.85%		80.65%		81.18%		74.26%	

ensembles with overall prediction rates of 81.80 percent and 80.21 percent, respectively (see Table 3.4). Even though the prediction results are slightly better than those of the individual models, ensembles are known to produce more robust prediction systems compared to a single-best prediction model (more on this can be found in Chapter 4).

In addition to assessing the prediction accuracy for each model type, a sensitivity analysis was also conducted using the developed prediction models to identify the relative importance of the independent variables (i.e., the predictors). In realizing the overall sensitivity analysis results, each of the four individual model types generated its own sensitivity measures, ranking all independent variables in a prioritized list. As expected, each model type generated slightly different sensitivity rankings of the independent variables. After collecting all four sets of sensitivity numbers, the sensitivity numbers are normalized and aggregated and plotted in a horizontal bar chart (see Figure 3.6).

The Conclusions

The study showed that, given sufficient data with the proper variables, data mining methods are capable of predicting freshmen student attrition with approximately 80 percent accuracy. Results also showed that, regardless of the prediction model employed, the balanced data set (compared to unbalanced/original data set) produced better prediction models for identifying the students who are likely to drop out of the college prior to their sophomore year. Among the four individual prediction models used in this study, support vector machines performed the best, followed by decision trees, neural networks, and logistic regression. From the usability standpoint, despite the fact that support vector machines showed better prediction results, one might choose to use decision trees, because compared to support vector machines and neural networks, they portray a more transparent model structure. Decision trees

TABLE 3.4 Prediction Results for the Three Ensemble Models

	Boosting (boosted trees)		Bagging (random forest)		Information Fusion (weighted average)	
	No	Yes	No	Yes	No	Yes
No	2,242	375	2,327	362	2,335	351
Yes	848	2,715	763	2,728	755	2,739
SUM	3,090	3,090	3,090	3,090	3,090	3,090
Per-class accuracy	72.56%	87.86%	75.31%	88.28%	75.57%	88.64%
Overall accuracy	80.21%		81.80%		82.10%	

(*Continued*)

Application Case 3.2 (Continued)

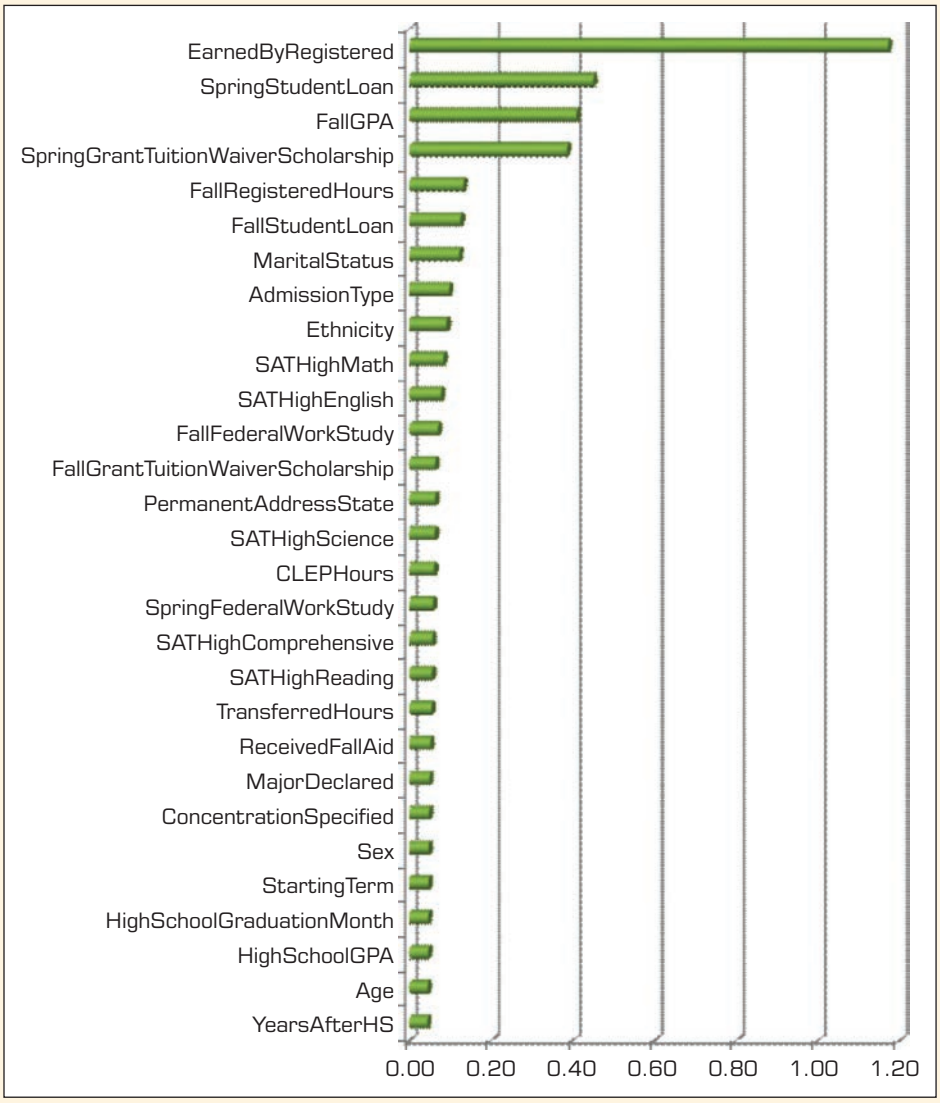


FIGURE 3.6 **Sensitivity-Analysis-Based Variable Importance Results.**

explicitly show the reasoning process of different predictions, providing a justification for a specific outcome, whereas support vector machines and artificial neural networks are mathematical models that do not provide such a transparent view of "how they do what they do."

Questions for Case 3.2

1. What is student attrition, and why is it an important problem in higher education?
2. What were the traditional methods to deal with the attrition problem?
3. List and discuss the data-related challenges within the context of this case study.
4. What was the proposed solution? What were the results?

Sources: D. Thammasiri, D. Delen, P. Meesad, & N. Kasap, "A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition," *Expert Systems with Applications, 41*(2), 2014, pp. 321–330; D. Delen, "A Comparative Analysis of Machine Learning Techniques for Student Retention Management," *Decision Support Systems, 49*(4), 2010, pp. 498–506, and "Predicting Student Attrition with Data Mining Methods," *Journal of College Student Retention 13*(1), 2011, pp. 17–35.

SECTION 3.4 REVIEW QUESTIONS

1. Why are the original/raw data not readily usable by analytics tasks?
2. What are the main data preprocessing steps?
3. What does it mean to clean/scrub the data? What activities are performed in this phase?
4. Why do we need data transformation? What are the commonly used data transformation tasks?
5. Data reduction can be applied to rows (sampling) and/or columns (variable selection). Which is more challenging?

3.5 STATISTICAL MODELING FOR BUSINESS ANALYTICS

Because of the increasing popularity of business analytics, the traditional statistical methods and underlying techniques are also regaining their attractiveness as enabling tools to support evidence-based managerial decision making. Not only are they regaining attention and admiration, but this time, they are attracting business users in addition to statisticians and analytics professionals.

Statistics (statistical methods and underlying techniques) is usually considered as part of descriptive analytics (see Figure 3.7). Some of the statistical methods can also be considered as part of predictive analytics, such as discriminant analysis, multiple regression, logistic regression, and k-means clustering. As shown in Figure 3.7, descriptive analytics has two main branches: statistics and **online analytics processing (OLAP)**. OLAP is the term used for analyzing, characterizing, and summarizing structured data stored in organizational databases (often stored in a data warehouse or in a data mart) using cubes (i.e., multidimensional data structures that are created to extract a subset of data values to answer a specific business question). The OLAP branch of descriptive analytics has also been called *business intelligence*. Statistics, on the other hand, helps to characterize the data, either one variable at a time or multivariable, all together using either descriptive or inferential methods.

Statistics—a collection of mathematical techniques to characterize and interpret data—has been around for a very long time. Many methods and techniques have been developed to address the needs of the end users and the unique characteristics of the data being analyzed. Generally speaking, at the highest level, statistical methods can be

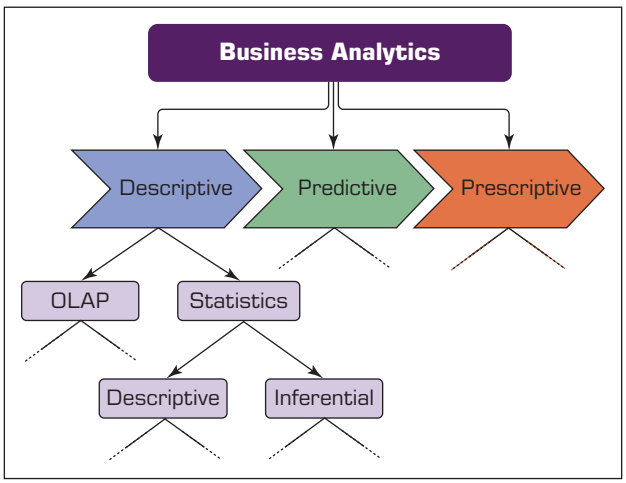


FIGURE 3.7 Relationship between Statistics and Descriptive Analytics.

classified as either descriptive or inferential. The main difference between descriptive and inferential statistics is the data used in these methods—whereas **descriptive statistics** is all about describing the sample data on hand, **inferential statistics** is about drawing inferences or conclusions about the characteristics of the population. In this section, we briefly describe descriptive statistics (because of the fact that it lays the foundation for, and is the integral part of, descriptive analytics), and in the following section we cover regression (both linear and logistic regression) as part of inferential statistics.

Descriptive Statistics for Descriptive Analytics

Descriptive statistics, as the name implies, describes the basic characteristics of the data at hand, often one variable at a time. Using formulas and numerical aggregations, descriptive statistics summarizes the data in such a way that often meaningful and easily understandable patterns emerge from the study. Although it is very useful in data analytics and very popular among the statistical methods, descriptive statistics does not allow making conclusions (or inferences) beyond the sample of the data being analyzed. That is, it is simply a nice way to characterize and describe the data on hand without making conclusions (inferences or extrapolations) regarding the population of related hypotheses we might have in mind.

In business analytics, descriptive statistics plays a critical role—it allows us to understand and explain/present our data in a meaningful manner using aggregated numbers, data tables, or charts/graphs. In essence, descriptive statistics helps us convert our numbers and symbols into meaningful representations for anyone to understand and use. Such an understanding helps not only business users in their decision-making processes but also analytics professionals and data scientists to characterize and validate the data for other more sophisticated analytics tasks. Descriptive statistics allows analysts to identify data concertation, unusually large or small values (i.e., outliers), and unexpectedly distributed data values for numeric variables. Therefore, the methods in descriptive statistics can be classified as either measures for central tendency or measures of dispersion. In the following section, we use a simple description and mathematical formulation/representation of these measures. In mathematical representation, we will use $x_1, x_2, \ldots, x_n$ to represent individual values (observations) of the variable (measure) that we are interested in characterizing.

Measures of Centrality Tendency (Also Called *Measures of Location or Centrality*)

Measures of centrality are the mathematical methods by which we estimate or describe central positioning of a given variable of interest. A measure of central tendency is a single numerical value that aims to describe a set of data by simply identifying or estimating the central position within the data. The mean (often called the *arithmetic mean* or the *simple average*) is the most commonly used measure of central tendency. In addition to mean, you could also see median or mode being used to describe the centrality of a given variable. Although, the mean, median, and mode are all valid measures of central tendency, under different circumstances, one of these measures of centrality becomes more appropriate than the others. What follows are short descriptions of these measures, including how to calculate them mathematically and pointers on the circumstances in which they are the most appropriate measure to use.

Arithmetic Mean

The **arithmetic mean** (or simply *mean* or *average*) is the sum of all the values/observations divided by the number of observations in the data set. It is by far the most popular

and most commonly used measure of central tendency. It is used with continuous or discrete numeric data. For a given variable x, if we happen to have n values/observations $(x_1, x_2, \ldots, x_n)$, we can write the arithmetic mean of the data sample ($\bar{x}$, pronounced as x-bar) as follows:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean has several unique characteristics. For instance, the sum of the absolute deviations (differences between the mean and the observations) above the mean is the same as the sum of the deviations below the mean, balancing the values on either side of it. That said, it does not suggest, however, that half the observations are above and the other half are below the mean (a common misconception among those who do not know basic statistics). Also, the mean is unique for every data set and is meaningful and calculable for both interval- and ratio-type numeric data. One major downside is that the mean can be affected by outliers (observations that are considerably larger or smaller than the rest of the data points). Outliers can pull the mean toward their direction and, hence, bias the centrality representation. Therefore, if there are outliers or if the data are erratically dispersed and skewed, one should either avoid using the mean as the measure of centrality or augment it with other central tendency measures, such as median and mode.

Median

The **median** is the measure of center value in a given data set. It is the number in the middle of a given set of data that has been arranged/sorted in order of magnitude (either ascending or descending). If the number of observations is an odd number, identifying the median is very easy—just sort the observations based on their values and pick the value right in the middle. If the number of observations is an even number, identify the two middle values, and then take the simple average of these two values. The median is meaningful and calculable for ratio, interval, and ordinal data types. Once determined, one-half of the data points in the data is above and the other half is below the median. In contrary to the mean, the median is not affected by outliers or skewed data.

Mode

The **mode** is the observation that occurs most frequently (the most frequent value in our data set). On a histogram, it represents the highest bar in a bar chart, and, hence, it can be considered as the most popular option/value. The mode is most useful for data sets that contain a relatively small number of unique values. That is, it could be useless if the data have too many unique values (as is the case in many engineering measurements that capture high precision with a large number of decimal places), rendering each value having either one or a very small number representing its frequency. Although it is a useful measure (especially for nominal data), mode is not a very good representation of centrality, and therefore, it should not be used as the only measure of central tendency for a given data set.

In summary, which central tendency measure is the best? Although there is not a clear answer to this question, here are a few hints—use the mean when the data are not prone to outliers and there is no significant level of skewness; use the median when the data have outliers and/or it is ordinal in nature; use the mode when the data are nominal.

Perhaps the best practice is to use all three together so that the central tendency of the data set can be captured and represented from three perspectives. Mostly because "average" is a very familiar and highly used concept to everyone in regular daily activities, managers (as well as some scientists and journalists) often use the centrality measures (especially mean) inappropriately when other statistical information should be considered along with the centrality. It is a better practice to present descriptive statistics as a package—a combination of centrality and dispersion measures—as opposed to a single measure such as mean.

Measures of Dispersion (Also Called *Measures of Spread* or *Decentrality*)

Measures of **dispersion** are the mathematical methods used to estimate or describe the degree of variation in a given variable of interest. They represent the numerical spread (compactness or lack thereof) of a given data set. To describe this dispersion, a number of statistical measures are developed; the most notable ones are range, variance, and standard deviation (and also quartiles and absolute deviation). One of the main reasons why the measures of dispersion/spread of data values are important is the fact that they give us a framework within which we can judge the central tendency—give us the indication of how well the mean (or other centrality measures) represents the sample data. If the dispersion of values in the data set is large, the mean is not deemed to be a very good representation of the data. This is because a large dispersion measure indicates large differences between individual scores. Also, in research, it is often perceived as a positive sign to see a small variation within each data sample, as it may indicate homogeneity, similarity, and robustness within the collected data.

Range

The **range** is perhaps the simplest measure of dispersion. It is the difference between the largest and the smallest values in a given data set (i.e., variables). So we calculate range by simply identifying the smallest value in the data set (minimum), identifying the largest value in the data set (maximum), and calculating the difference between them (range = maximum − minimum).

Variance

A more comprehensive and sophisticated measure of dispersion is the **variance**. It is a method used to calculate the deviation of all data points in a given data set from the mean. The larger the variance, the more the data are spread out from the mean and the more variability one can observe in the data sample. To prevent the offsetting of negative and positive differences, the variance takes into account the square of the distances from the mean. The formula for a data sample can be written as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

where n is the number of samples, $\bar{x}$ is the mean of the sample, and x_i is the i^{th} value in the data set. The larger values of variance indicate more dispersion, whereas smaller values indicate compression in the overall data set. Because the differences are squared, larger deviations from the mean contribute significantly to the value of variance. Again, because the differences are squared, the numbers that represent deviation/variance become somewhat meaningless (as opposed to a dollar difference, here you are given a squared dollar difference). Therefore, instead of variance, in

many business applications, we use a more meaningful dispersion measure, called *standard deviation*.

Standard Deviation

The **standard deviation** is also a measure of the spread of values within a set of data. The standard deviation is calculated by simply taking the square root of the variations. The following formula shows the calculation of standard deviation from a given sample of data points.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

Mean Absolute Deviation

In addition to variance and standard deviation, sometimes we also use **mean absolute deviation** to measure dispersion in a data set. It is a simpler way to calculate the overall deviation from the mean. Specifically, the mean absolute deviation is calculated by measuring the absolute values of the differences between each data point and the mean and then summing them. This process provides a measure of spread without being specific about the data point being lower or higher than the mean. The following formula shows the calculation of the mean absolute deviation:

$$MAD = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$$

Quartiles and Interquartile Range

Quartiles help us identify spread within a subset of the data. A **quartile** is a quarter of the number of data points given in a data set. Quartiles are determined by first sorting the data and then splitting the sorted data into four disjoint smaller data sets. Quartiles are a useful measure of dispersion because they are much less affected by outliers or a skewness in the data set than the equivalent measures in the whole data set. Quartiles are often reported along with the median as the best choice of measure of dispersion and central tendency, respectively, when dealing with skewed and/or data with outliers. A common way of expressing quartiles is as an interquartile range, which describes the difference between the third quartile (Q3) and the first quartile (Q1), telling us about the range of the middle half of the scores in the distribution. The quartile-driven descriptive measures (both centrality and dispersion) are best explained with a popular plot called a *box-and-whiskers plot* (or *box plot*).

Box-and-Whiskers Plot

The **box-and-whiskers plot** (or simply a **box plot**) is a graphical illustration of several descriptive statistics about a given data set. They can be either horizontal or vertical, but vertical is the most common representation, especially in modern-day analytics software products. It is known to be first created and presented by John W. Tukey in 1969. Box plot is often used to illustrate both centrality and dispersion of a given data set (i.e., the distribution of the sample data) in an easy-to-understand graphical notation. Figure 3.8 shows two box plots side by side, sharing the same *y*-axis. As shown therein, a single chart can have one or more box plots for visual comparison purposes. In such cases, the *y*-axis would be the common measure of magnitude (the numerical value of the

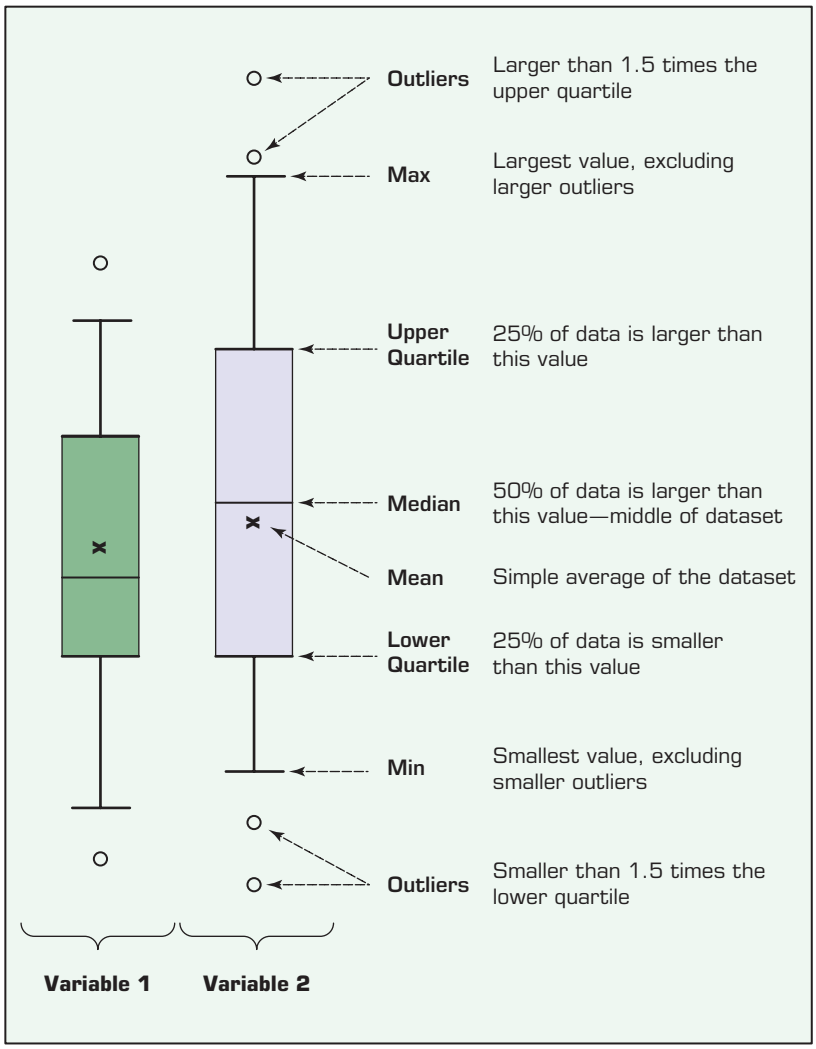


FIGURE 3.8 **Understanding the Specifics about Box-and-Whiskers Plots.**

variable), with the *x*-axis showing different classes/subsets such as different time dimensions (e.g., descriptive statistics for annual Medicare expenses in 2015 versus 2016) or different categories (e.g., descriptive statistics for marketing expenses versus total sales).

Although historically speaking, the box plot has not been used widely and often enough (especially in areas outside of statistics), with the emerging popularity of business analytics, it is gaining fame in less technical areas of the business world. Its information richness and ease of understanding are largely to credit for its recent popularity.

The box plot shows the **centrality** (median and sometimes also mean) as well as the dispersion (the density of the data within the middle half—drawn as a box between the first and third quartiles), the minimum and maximum ranges (shown as extended lines from the box, looking like whiskers, that are calculated as 1.5 times the upper or lower end of the quartile box), and the outliers that are larger than the limits of the whiskers. A box plot also shows whether the data are symmetrically distributed with respect to the mean or sway one way or another. The relative position of the median versus mean and the lengths of the whiskers on both side of the box give a good indication of the potential skewness in the data.

Shape of a Distribution

Although not as common as the centrality and dispersion, the shape of the data distribution is also a useful measure for the descriptive statistics. Before delving into the shape of the distribution, we first need to define the distribution itself. Simply put, *distribution* is the frequency of data points counted and plotted over a small number of class labels or numerical ranges (i.e., bins). In a graphical illustration of distribution, the *y*-axis shows the frequency (count or percentage), and the *x*-axis shows the individual classes or bins in a rank-ordered fashion. A very well-known distribution is called *normal distribution*, which is perfectly symmetric on both sides of the mean and has numerous well-founded mathematical properties that make it a very useful tool for research and practice. As the dispersion of a data set increases, so does the standard deviation, and the shape of the distribution looks wider. A graphic illustration of the relationship between dispersion and distribution shape (in the context of normal distribution) is shown in Figure 3.9.

There are two commonly used measures to calculate the shape characteristics of a distribution: skewness and kurtosis. A histogram (frequency plot) is often used to visually illustrate both skewness and kurtosis.

Skewness is a measure of asymmetry (sway) in a distribution of the data that portrays a unimodal structure—only one peak exists in the distribution of the data. Because normal distribution is a perfectly symmetric unimodal distribution, it does not have

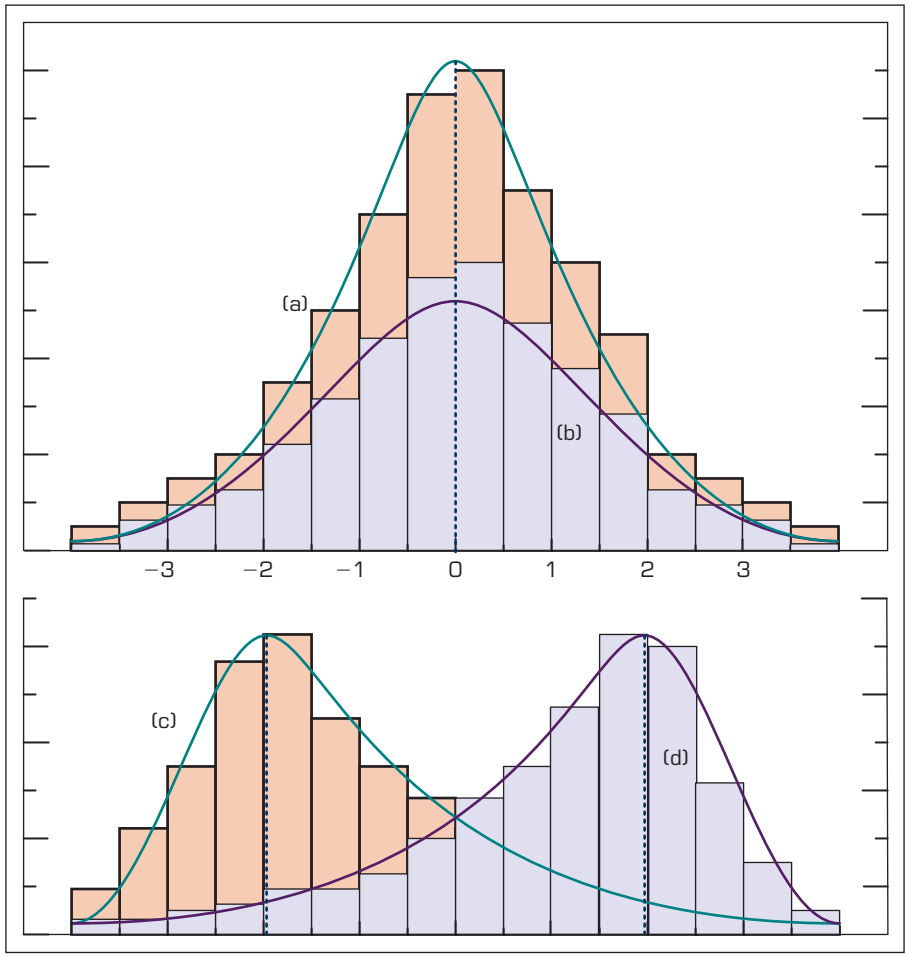


FIGURE 3.9 **Relationship between Dispersion and Distribution Shape Properties.**

skewness; that is, its skewness measure (i.e., the value of the coefficient of skewness) is equal to zero. The skewness measure/value can be either positive or negative. If the distribution sways left (i.e., the tail is on the right side and the mean is smaller than median), then it produces a positive skewness measure; if the distribution sways right (i.e., the tail is on the left side and the mean is larger than median), then it produces a negative skewness measure. In Figure 3.9, (c) represents a positively skewed distribution whereas (d) represents a negatively skewed distribution. In the same figure, both (a) and (b) represent perfect symmetry and hence zero measure for skewness.

$$Skewness = S = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

where s is the standard deviation and n is the number of samples.

Kurtosis is another measure to use in characterizing the shape of a unimodal distribution. As opposed to the sway in shape, kurtosis focuses more on characterizing the peak/tall/skinny nature of the distribution. Specifically, kurtosis measures the degree to which a distribution is more or less peaked than a normal distribution. Whereas a positive kurtosis indicates a relatively peaked/tall distribution, a negative kurtosis indicates a relatively flat/short distribution. As a reference point, a normal distribution has a kurtosis of 3. The formula for kurtosis can be written as

$$Kurtosis = K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

Descriptive statistics (as well as inferential statistics) can easily be calculated using commercially viable statistical software packages (e.g., SAS, SPSS, Minitab, JMP, Statistica) or free/open source tools (e.g., R). Perhaps the most convenient way to calculate descriptive and some of the inferential statistics is to use Excel. Technology Insights 3.1 describes in detail how to use Microsoft Excel to calculate descriptive statistics.

TECHNOLOGY INSIGHTS 3.1 How to Calculate Descriptive Statistics in Microsoft Excel

Excel, arguably the most popular data analysis tool in the world, can easily be used for descriptive statistics. Although the base configuration of Excel does not seem to have the statistics function readily available for end users, those functions come with the Excel installation and can be activated (turned on) with only a few mouse clicks. Figure 3.10 shows how these statistics functions (as part of the Analysis ToolPak) can be activated in Microsoft Excel 2016.

Once activated, the *Analysis ToolPak* will appear in the *Data* menu option under the name of *Data Analysis*. When you click on Data Analysis in the Analysis group under the Data tab in the Excel menu bar, you will see Descriptive Statistics as one of the options within the list of data analysis tools (see Figure 3.11, steps 1, 2); click on OK, and the Descriptive Statistics dialog box will appear (see the middle of Figure 3.11). In this dialog box, you need to enter the range of the data, which can be one or more numerical columns, along with the preference check boxes, and click OK (see Figure 3.11, steps 3, 4). If the selection includes more than one numeric column, the tool treats each column as a separate data set and provides descriptive statistics for each column separately.

As a simple example, we selected two columns (labeled as Expense and Demand) and executed the Descriptive Statistics option. The bottom section of Figure 3.11 shows the output created by Excel. As can be seen, Excel produced all descriptive statistics that are covered in the previous section and added a few more to the list. In Excel 2016, it is also very easy (a few

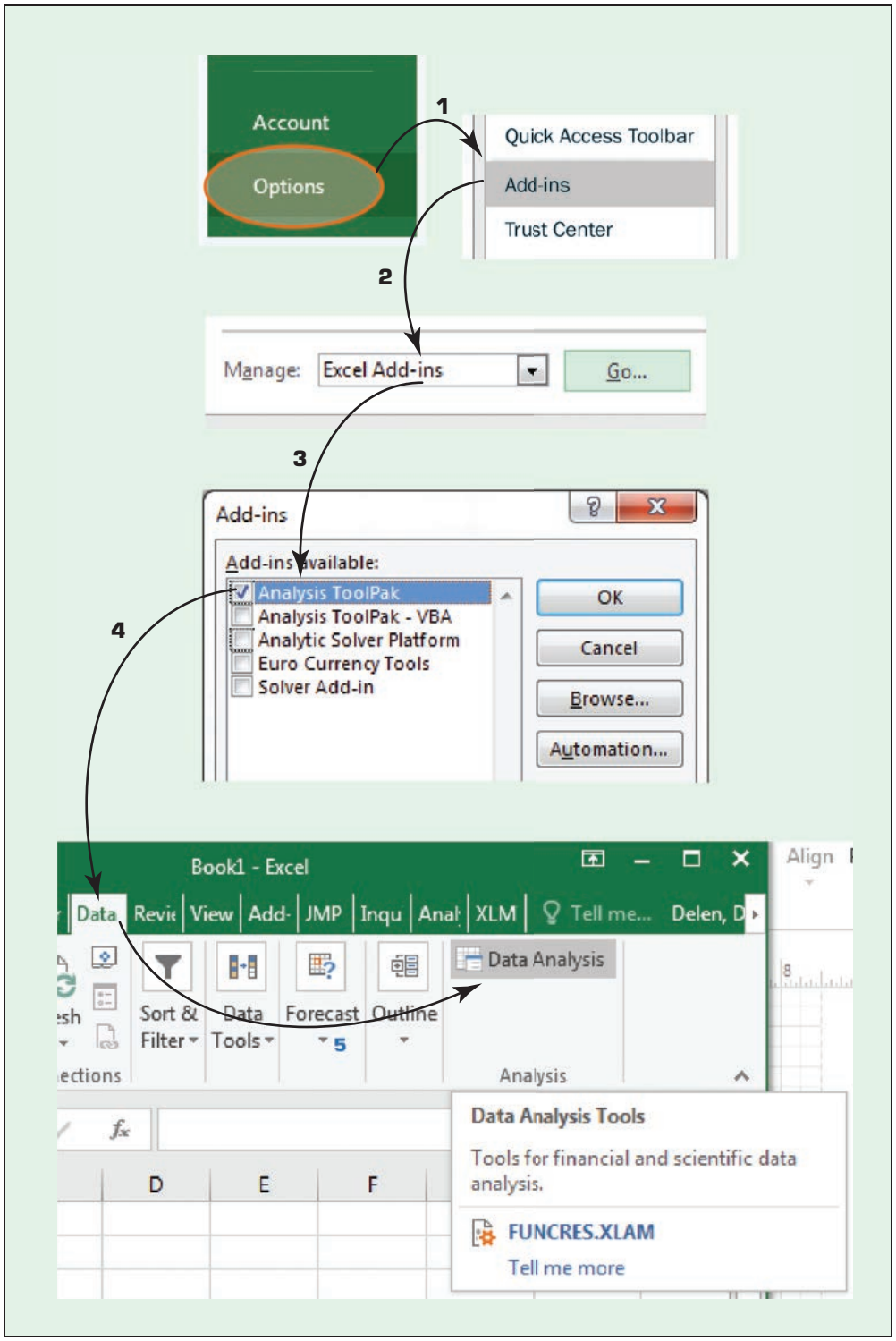


FIGURE 3.10 **Activating Statistics Function in Excel 2016.**

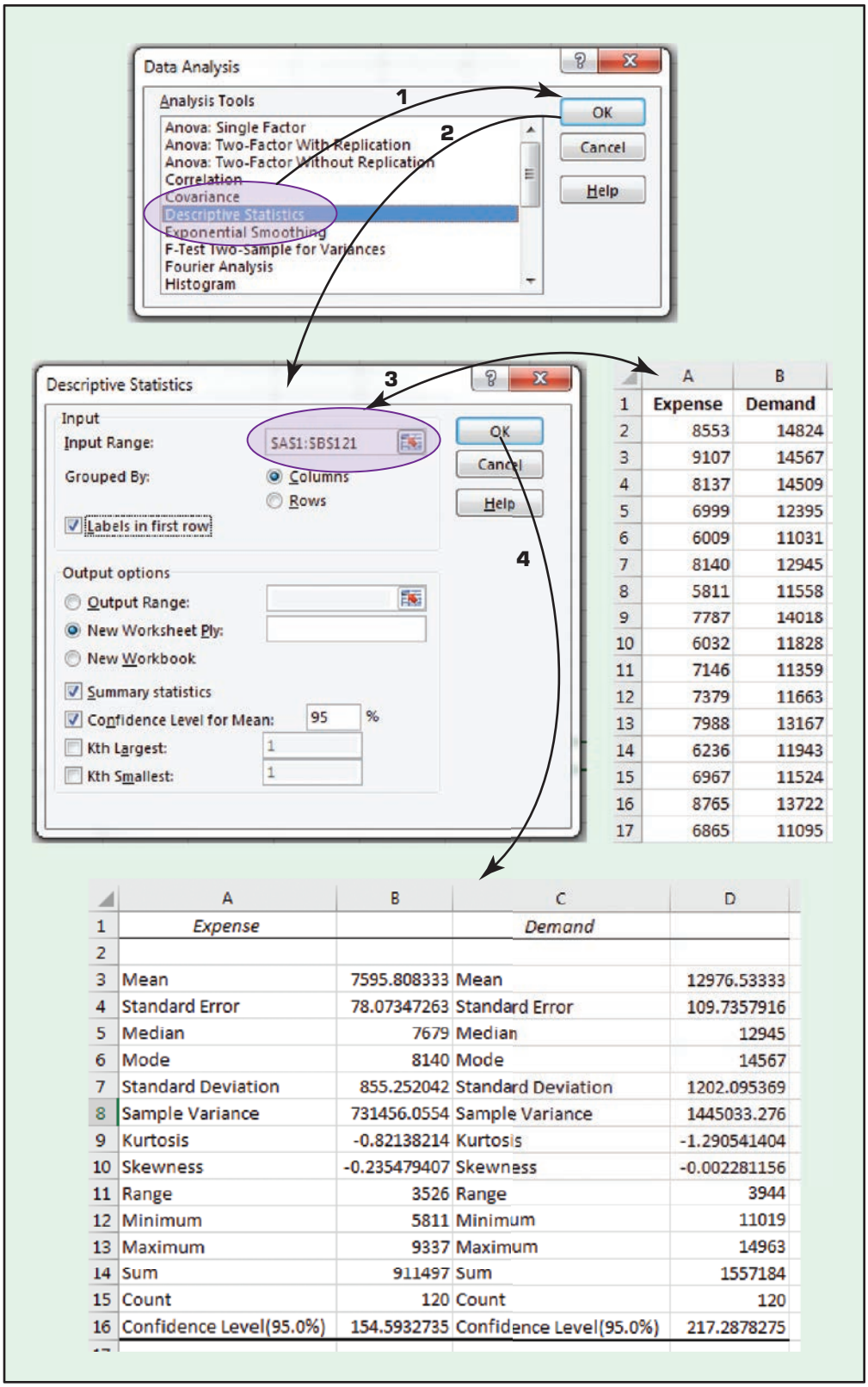


FIGURE 3.11 Obtaining Descriptive Statistics in Excel.

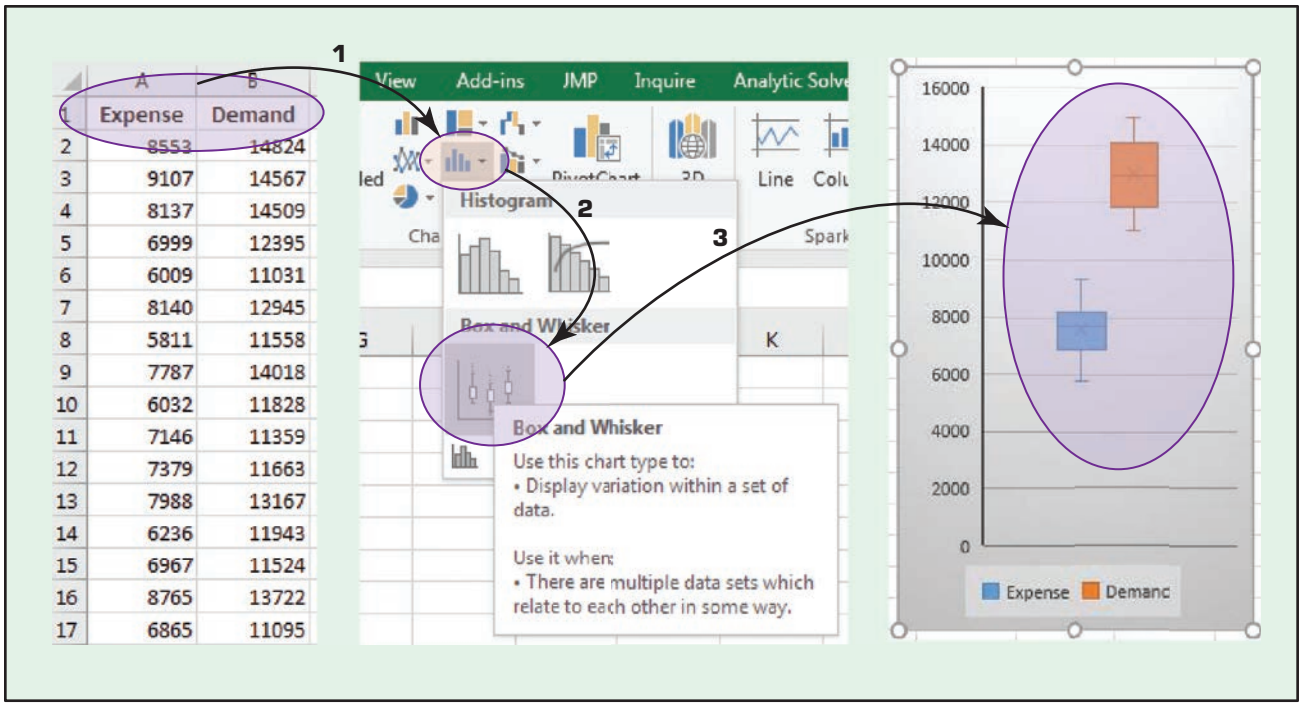


FIGURE 3.12 **Creating a Box-and-Whiskers Plot in Excel 2016.**

mouse clicks) to create a box-and-whiskers plot. Figure 3.12 shows the simple three-step process of creating a box-and-whiskers plot in Excel.

Although Analysis ToolPak is a very useful tool in Excel, one should be aware of an important point related to the results that it generates, which have a different behavior than other ordinary Excel functions: Although Excel functions dynamically change as the underlying data in the spreadsheet are changed, the results generated by the Analysis ToolPak do not. For example, if you change the values in either or both of these columns, the Descriptive Statistics results produced by the Analysis ToolPak will stay the same. However, the same is not true for ordinary Excel functions. If you were to calculate the mean value of a given column (using "=AVERAGE(A1:A121)") and then change the values within the data range, the mean value would automatically change. In summary, the results produced by Analysis ToolPak do not have a dynamic link to the underlying data, and if the data change, the analysis needs to be redone using the dialog box.

Successful applications of data analytics cover a wide range of business and organizational settings, addressing problems once thought unsolvable. Application Case 3.3 is an excellent illustration of those success stories in which a small municipality administration adopted a data analytics approach to intelligently detect and solve problems by continuously analyzing demand and consumption patterns.

SECTION 3.5 REVIEW QUESTIONS

1. What is the relationship between statistics and business analytics?
2. What are the main differences between descriptive and inferential statistics?
3. List and briefly define the central tendency measures of descriptive statistics.
4. List and briefly define the dispersion measures of descriptive statistics.
5. What is a box-and-whiskers plot? What types of statistical information does it represent?
6. What are the two most commonly used shape characteristics to describe a data distribution?

Application Case 3.3

Town of Cary Uses Analytics to Analyze Data from Sensors, Assess Demand, and Detect Problems

A leaky faucet. A malfunctioning dishwasher. A cracked sprinkler head. These are more than just a headache for a home owner or business to fix. They can be costly, unpredictable, and, unfortunately, hard to pinpoint. Through a combination of wireless water meters and a data-analytics-driven, customer-accessible portal, the Town of Cary, North Carolina, is making it much easier to find and fix water loss issues. In the process, the town has gained a big-picture view of water usage critical to planning future water plant expansions and promoting targeted conservation efforts.

When the town of Cary installed wireless meters for 60,000 customers in 2010, it knew the new technology wouldn't just save money by eliminating manual monthly readings; the town also realized it would get more accurate and timely information about water consumption. The Aquastar wireless system reads meters once an hour—that is 8,760 data points per customer each year instead of 12 monthly readings. The data had tremendous potential if they could be easily consumed.

"Monthly readings are like having a gallon of water's worth of data. Hourly meter readings are more like an Olympic-size pool of data," says Karen Mills, finance director for Cary. "SAS helps us manage the volume of that data nicely." In fact, the solution enables the town to analyze half a billion data points on water usage and make them available to and easily consumable by all customers.

The ability to visually look at data by household or commercial customer by the hour has led to some very practical applications:

- The town can notify customers of potential leaks within days.
- Customers can set alerts that notify them within hours if there is a spike in water usage.
- Customers can track their water usage online, helping them to be more proactive in conserving water.

Through the online portal, one business in the town saw a spike in water consumption on weekends when employees are away. This seemed odd, and the unusual reading helped the company learn that a commercial dishwasher was malfunctioning, running continuously over weekends. Without the wireless water-meter data and the customer-accessible portal, this problem could have gone unnoticed, continuing to waste water and money.

The town has a much more accurate picture of daily water usage per person, critical for planning future water plant expansions. Perhaps the most interesting perk is that the town was able to verify a hunch that has far-reaching cost ramifications: Cary residents are very economical in their use of water. "We calculate that with modern high-efficiency appliances, indoor water use could be as low as 35 gallons per person per day. Cary residents average 45 gallons, which is still phenomenally low," explains town Water Resource Manager Leila Goodwin. Why is this important? The town was spending money to encourage water efficiency—rebates on low-flow toilets or discounts on rain barrels. Now it can take a more targeted approach, helping specific consumers understand and manage both their indoor and outdoor water use.

SAS was critical not just for enabling residents to understand their water use but also working behind the scenes to link two disparate databases. "We have a billing database and the meter-reading database. We needed to bring that together and make it presentable," Mills says.

The town estimates that by just removing the need for manual readings, the Aquastar system will save more than $10 million above the cost of the project. But the analytics component could provide even bigger savings. Already, both the town and individual citizens have saved money by catching water leaks early. As Cary continues to plan its future infrastructure needs, having accurate information on water usage will help it invest in the right amount of infrastructure at the right time. In addition, understanding water usage will help the town if it experiences something detrimental like a drought.

"We went through a drought in 2007," says Goodwin. "If we go through another, we have a plan in place to use Aquastar data to see exactly how much water we are using on a day-by-day basis and communicate with customers. We can show 'here's what's happening, and here is how much you can use because our supply is low.' Hopefully, we'll never have to use it, but we're prepared."

Questions for Case 3.3

1. What were the challenges the Town of Cary was facing?
2. What was the proposed solution?
3. What were the results?
4. What other problems and data analytics solutions do you foresee for towns like Cary?

Source: "Municipality Puts Wireless Water Meter-Reading Data To Work (SAS® Analytics)—The Town of Cary, North Carolina Uses SAS Analytics to Analyze Data from Wireless Water Meters, Assess Demand, Detect Problems and Engage Customers." Copyright © 2016 SAS Institute Inc., Cary, NC, USA. Reprinted with permission. All rights reserved.

3.6 REGRESSION MODELING FOR INFERENTIAL STATISTICS

Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics. Historically speaking, the roots of regression date back to the 1920s and 1930s, to the earlier work on inherited characteristics of sweet peas by Sir Francis Galton and subsequently by Karl Pearson. Since then, regression has become the statistical technique for characterization of relationships between explanatory (input) variable(s) and response (output) variable(s).

As popular as it is, regression essentially is a relatively simple statistical technique to model the dependence of a variable (response or output variable) on one (or more) explanatory (input) variables. Once identified, this relationship between the variables can be formally represented as a linear/additive function/equation. As is the case with many other modeling techniques, regression aims to capture the functional relationship between and among the characteristics of the real world and describe this relationship with a mathematical model, which can then be used to discover and understand the complexities of reality—explore and explain relationships or forecast future occurrences.

Regression can be used for one of two purposes: hypothesis testing—investigating potential relationships between different variables—and prediction/forecasting—estimating values of a response variable based on one or more explanatory variables. These two uses are not mutually exclusive. The explanatory power of regression is also the foundation of its predictive ability. In hypothesis testing (theory building), regression analysis can reveal the existence/strength and the directions of relationships between a number of explanatory variables (often represented with x_i) and the response variable (often represented with y). In prediction, regression identifies additive mathematical relationships (in the form of an equation) between one or more explanatory variables and a response variable. Once determined, this equation can be used to forecast the values of the response variable for a given set of values of the explanatory variables.

CORRELATION VERSUS REGRESSION Because regression analysis originated from correlation studies, and because both methods attempt to describe the association between two (or more) variables, these two terms are often confused by professionals and even by scientists. **Correlation** makes no a priori assumption of whether one variable is dependent on the other(s) and is not concerned with the relationship between variables; instead it gives an estimate on the degree of association between the variables. On the other hand, regression attempts to describe the dependence of a response variable on one (or more) explanatory variables where it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect. Also, although correlation is interested in the low-level relationships between two variables, regression is concerned with the relationships between all explanatory variables and the response variable.

SIMPLE VERSUS MULTIPLE REGRESSION If the regression equation is built between one response variable and one explanatory variable, then it is called *simple regression*. For instance, the regression equation built to predict/explain the relationship between the height of a person (explanatory variable) and the weight of a person (response variable) is a good example of simple regression. Multiple regression is the extension of simple regression when the explanatory variables are more than one. For instance, in the previous example, if we were to include not only the height of the person but also other personal characteristics (e.g., BMI, gender, ethnicity) to predict the person's weight, then we would be performing multiple regression analysis. In both cases, the relationship between the response variable and the explanatory variable(s) is linear and additive in nature. If the relationships are not linear, then we might want to use one of many other nonlinear regression methods to better capture the relationships between the input and output variables.

How Do We Develop the Linear Regression Model?

To understand the relationship between two variables, the simplest thing that one can do is to draw a scatter plot where the *y*-axis represents the values of the response variable and the *x*-axis represents the values of the explanatory variable (see Figure 3.13). A scatter plot would show the changes in the response variable as a function of the changes in the explanatory variable. In the case shown in Figure 3.13, there seems to be a positive relationship between the two; as the explanatory variable values increase, so does the response variable.

Simple regression analysis aims to find a mathematical representation of this relationship. In reality, it tries to find the signature of a straight line passing through right between the plotted dots (representing the observation/historical data) in such a way that it minimizes the distance between the dots and the line (the predicted values on the

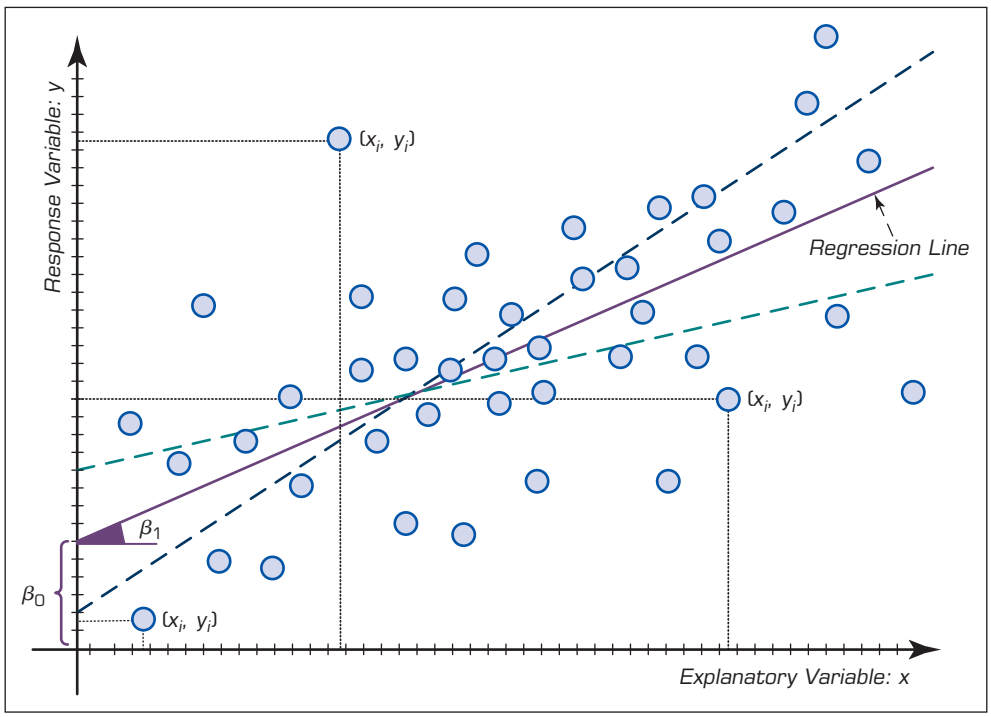


FIGURE 3.13 A Scatter Plot and a Linear Regression Line.

theoretical regression line). Even though there are several methods/algorithms proposed to identify the regression line, the one that is most commonly used is called the **ordinary least squares (OLS)** method. The OLS method aims to minimize the sum of squared residuals (squared vertical distances between the observation and the regression point) and leads to a mathematical expression for the estimated value of the regression line (which are known as β parameters). For simple **linear regression**, the aforementioned relationship between the response variable (y) and the explanatory variable(s) (x) can be shown as a simple equation as follows:

$$y = \beta_0 + \beta_1 x$$

In this equation, β_0 is called the intercept, and β_1 is called the slope. Once OLS determines the values of these two coefficients, the simple equation can be used to forecast the values of y for given values of x. The sign and the value of β_1 also reveal the direction and the strengths of relationship between the two variables.

If the model is of a multiple linear regression type, then there would be more coefficients to be determined, one for each additional explanatory variable. As the following formula shows, the additional explanatory variable would be multiplied with the new β_i coefficients and summed together to establish a linear additive representation of the response variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdot + \beta_n x_n$$

How Do We Know If the Model Is Good Enough?

Because of a variety of reasons, sometimes models as representations of the reality do not prove to be good. Regardless of the number of explanatory variables included, there is always a possibility of not having a good model, and therefore the linear regression model needs to be assessed for its fit (the degree to which it represents the response variable). In the simplest sense, a well-fitting regression model results in predicted values close to the observed data values. For the numerical assessment, three statistical measures are often used in evaluating the fit of a regression model: R^2(R − squared), the overall F-test, and the root mean square error (RMSE). All three of these measures are based on the sums of the square errors (how far the data are from the mean and how far the data are from the model's predicted values). Different combinations of these two values provide different information about how the regression model compares to the mean model.

Of the three, R^2 has the most useful and understandable meaning because of its intuitive scale. The value of R^2 ranges from 0 to 1 (corresponding to the amount of variability explained in percentage) with 0 indicating that the relationship and the prediction power of the proposed model is not good, and 1 indicating that the proposed model is a perfect fit that produces exact predictions (which is almost never the case). The good R^2 values would usually come close to one, and the closeness is a matter of the phenomenon being modeled—whereas an R^2 value of 0.3 for a linear regression model in social sciences can be considered good enough, an R^2 value of 0.7 in engineering might be considered as not a good enough fit. The improvement in the regression model can be achieved by adding more explanatory variables or using different data transformation techniques, which would result in comparative increases in an R^2 value. Figure 3.14 shows the process flow of developing regression models. As can be seen in the process flow, the model development task is followed by the model assessment task in which not only is the fit of the model assessed, but because of restrictive assumptions with which the linear models have to comply, the validity of the model also needs to be put under the microscope.

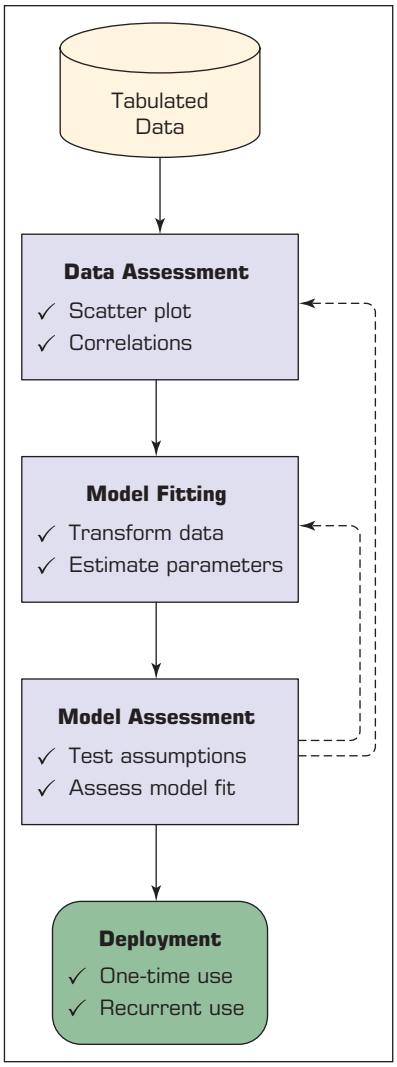


FIGURE 3.14 A Process Flow for Developing Regression Models.

What Are the Most Important Assumptions in Linear Regression?

Even though they are still the choice of many for data analyses (both for explanatory and for predictive modeling purposes), linear regression models suffer from several highly restrictive assumptions. The validity of the linear model built depends on its ability to comply with these assumptions. Here are the most commonly pronounced assumptions:

1. **Linearity.** This assumption states that the relationship between the response variable and the explanatory variables is linear. That is, the expected value of the response variable is a straight-line function of each explanatory variable while holding all other explanatory variables fixed. Also, the slope of the line does not depend on the values of the other variables. It also implies that the effects of different explanatory variables on the expected value of the response variable are additive in nature.
2. **Independence (of errors).** This assumption states that the errors of the response variable are uncorrelated with each other. This independence of the errors is weaker

than actual statistical independence, which is a stronger condition and is often not needed for linear regression analysis.

3. **Normality (of errors).** This assumption states that the errors of the response variable are normally distributed. That is, they are supposed to be totally random and should not represent any nonrandom patterns.
4. **Constant variance (of errors).** This assumption, also called *homoscedasticity*, states that the response variables have the same variance in their error regardless of the values of the explanatory variables. In practice, this assumption is invalid if the response variable varies over a wide enough range/scale.
5. **Multicollinearity.** This assumption states that the explanatory variables are not correlated (i.e., do not replicate the same but provide a different perspective of the information needed for the model). Multicollinearity can be triggered by having two or more perfectly correlated explanatory variables presented to the model (e.g., if the same explanatory variable is mistakenly included in the model twice, one with a slight transformation of the same variable). A correlation-based data assessment usually catches this error.

There are statistical techniques developed to identify the violation of these assumptions and techniques to mitigate them. The most important part for a modeler is to be aware of their existence and to put in place the means to assess the models to make sure that they are compliant with the assumptions they are built on.

Logistic Regression

Logistic regression is a very popular, statistically sound, probability-based classification algorithm that employs supervised **learning**. It was developed in the 1940s as a complement to linear regression and linear discriminant analysis methods. It has been used extensively in numerous disciplines, including the medical and social sciences fields. Logistic regression is similar to linear regression in that it also aims to regress to a mathematical function that explains the relationship between the response variable and the explanatory variables using a sample of past observations (training data). Logistic regression differs from linear regression with one major point: its output (response variable) is a class as opposed to a numerical variable. That is, whereas linear regression is used to estimate a continuous numerical variable, logistic regression is used to classify a categorical variable. Even though the original form of logistic regression was developed for a binary output variable (e.g., 1/0, yes/no, pass/fail, accept/reject), the present-day modified version is capable of predicting multiclass output variables (i.e., multinomial logistic regression). If there is only one predictor variable and one predicted variable, the method is called *simple logistic regression* (similar to calling linear regression models with only one independent variable *simple linear regression*).

In predictive analytics, logistic regression models are used to develop probabilistic models between one or more explanatory/predictor variables (which can be a mix of both continuous and categorical in nature) and a class/response variable (which can be binomial/binary or multinomial/multiclass). Unlike ordinary linear regression, logistic regression is used for predicting categorical (often binary) outcomes of the response variable—treating the response variable as the outcome of a Bernoulli trial. Therefore, logistic regression takes the natural logarithm of the odds of the response variable to create a continuous criterion as a transformed version of the response variable. Thus, the logit transformation is referred to as the *link function* in logistic regression—even though the response variable in logistic regression is categorical or binomial, the logit is the continuous criterion on which linear regression is conducted. Figure 3.15 shows a logistic

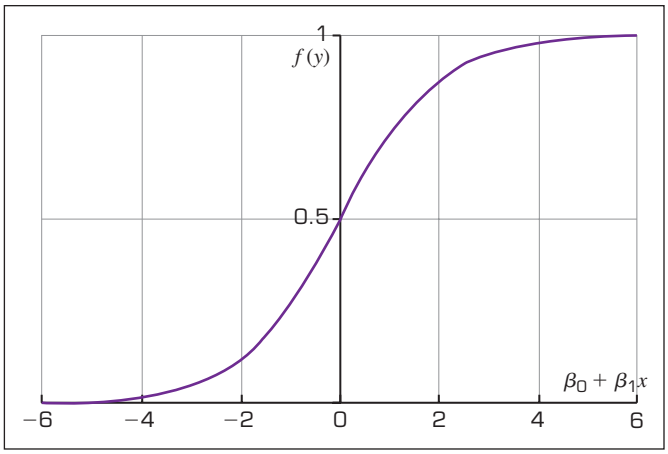


FIGURE 3.15 **The Logistic Function.**

regression function where the odds are represented in the *x-axis* (a linear function of the independent variables), whereas the probabilistic outcome is shown in the *y-axis* (i.e., response variable values change between 0 and 1).

The logistic function, $f(y)$ in Figure 3.15 is the core of logistic regression, which can take values only between 0 and 1. The following equation is a simple mathematical representation of this function:

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

The logistic regression coefficients (the βs) are usually estimated using the maximum likelihood estimation method. Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximizes the likelihood function, so an iterative process must be used instead. This process begins with a tentative starting solution, then revises the parameters slightly to see if the solution can be improved, and repeats this iterative revision until no improvement can be achieved or is very minimal, at which point the process is said to have completed/converged.

Sports analytics—use of data and statistical/analytics techniques to better manage sports teams/organizations—has been gaining tremendous popularity. Use of data-driven analytics techniques has become mainstream for not only professional teams but also college and amateur sports. Application Case 3.4 is an example of how existing and readily available public data sources can be used to predict college football bowl game outcomes using both classification and regression-type prediction models.

Time-Series Forecasting

Sometimes the variable that we are interested in (i.e., the response variable) might not have distinctly identifiable explanatory variables, or there might be too many of them in a highly complex relationship. In such cases, if the data are available in a desired format, a prediction model, the so-called time series, can be developed. A time series is a sequence of data points of the variable of interest, measured and represented at successive points in time spaced at uniform time intervals. Examples of time series include monthly rain volumes in a geographic area, the daily closing value of the stock market indexes, and

Application Case 3.4 Predicting NCAA Bowl Game Outcomes

Predicting the outcome of a college football game (or any sports game, for that matter) is an interesting and challenging problem. Therefore, challenge-seeking researchers from both academics and industry have spent a great deal of effort on forecasting the outcome of sporting events. Large amounts of historic data exist in different media outlets (often publicly available) regarding the structure and outcomes of sporting events in the form of a variety of numerically or symbolically represented factors that are assumed to contribute to those outcomes.

The end-of-season bowl games are very important to colleges in terms of both finance (bringing in millions of dollars of additional revenue) and reputation—for recruiting quality students and highly regarded high school athletes for their athletic programs (Freeman & Brewer, 2016). Teams that are selected to compete in a given bowl game split a purse, the size of which depends on the specific bowl (some bowls are more prestigious and have higher payouts for the two teams), and therefore securing an invitation to a bowl game is the main goal of any division I-A college football program. The decision makers of the bowl games are given the authority to select and invite bowl-eligible (a team that has six wins against its Division I-A opponents in that season) successful teams (as per the ratings and rankings) that will play in an exciting and competitive game, attract fans of both schools, and keep the remaining fans tuned in via a variety of media outlets for advertising.

In a recent data mining study, Delen et al. (2012) used eight years of bowl game data along with three popular data mining techniques (decision trees, neural networks, and support vector machines) to predict both the classification-type outcome of a game (win versus loss) and the regression-type outcome (projected point difference between the scores of the two opponents). What follows is a shorthand description of their study.

The Methodology

In this research, Delen and his colleagues followed a popular data mining methodology, *CRISP-DM* (Cross-Industry Standard Process for Data Mining), which is a six-step process. This popular methodology, which is covered in detail in Chapter 4, provided them with a systematic and structured way to conduct the underlying data mining study and hence improved the likelihood of obtaining accurate

(*Continued*)

Application Case 3.4 (Continued)

and reliable results. To objectively assess the prediction power of the different model types, they used a cross-validation methodology *k*-fold cross-validation. Details on *k*-fold cross-validation can be found in Chapter 4. Figure 3.16 graphically illustrates the methodology employed by the researchers.

Data Acquisition and Data Preprocessing

The sample data for this study are collected from a variety of sports databases available on the Web, including **jhowel.net**, **ESPN.com**, **Covers.com**, ncaa.org, and **rauzulusstreet.com**. The data set included 244 bowl games representing a complete set of eight seasons of college football bowl games played between 2002 and 2009. Delen et al. also included an out-of-sample data set (2010–2011 bowl games) for additional validation purposes. Exercising one of the popular data mining rules of thumb, they included as much relevant information in the model as possible. Therefore, after an in-depth variable identification and

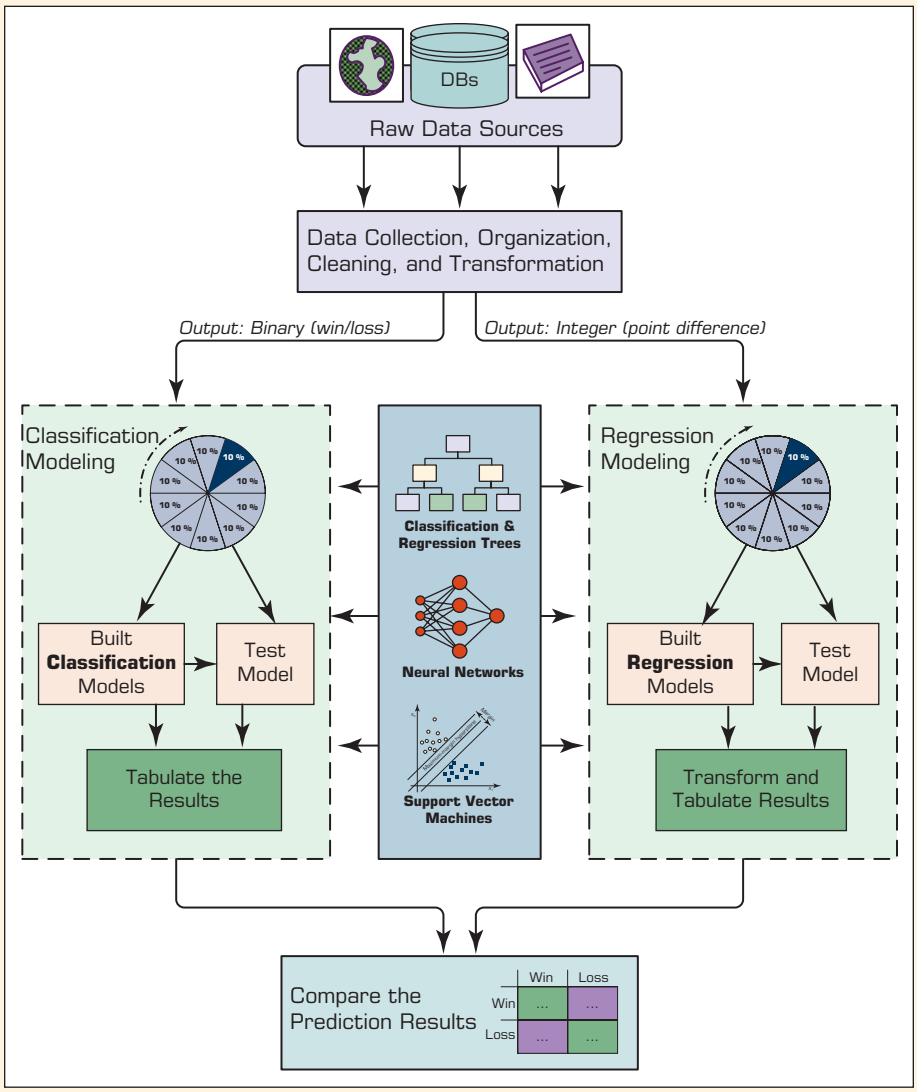


FIGURE 3.16 The Graphical Illustration of the Methodology Employed in the Study.

collection process, they ended up with a data set that included 36 variables, of which the first 6 were the identifying variables (i.e., name and the year of the bowl game, home and away team names, and their athletic conferences—see variables 1–6 in Table 3.5), followed by 28 input variables (which included variables delineating a team's seasonal statistics on offense and defense, game outcomes, team composition characteristics, athletic conference characteristics, and how they fared against the odds—see

TABLE 3.5 Description of Variables Used in the Study

No	Cat	Variable Name	Description
1	ID[1]	YEAR	Year of the bowl game
2	ID	BOWLGAME	Name of the bowl game
3	ID	HOMETEAM	Home team (as listed by the bowl organizers)
4	ID	AWAYTEAM	Away team (as listed by the bowl organizers)
5	ID	HOMECONFERENCE	Conference of the home team
6	ID	AWAYCONFERENCE	Conference of the away team
7	I1[2]	DEFPTPGM	Defensive points per game
8	I1	DEFRYDPGM	Defensive rush yards per game
9	I1	DEFYDPGM	Defensive yards per game
10	I1	PPG	Average number of points a given team scored per game
11	I1	PYDPGM	Average total pass yards per game
12	I1	RYDPGM	Team's average total rush yards per game
13	I1	YRDPGM	Average total offensive yards per game
14	I2	HMWIN%	Home winning percentage
15	I2	LAST7	How many games the team won out of their last 7 games
16	I2	MARGOVIC	Average margin of victory
17	I2	NCTW	Nonconference team winning percentage
18	I2	PREVAPP	Did the team appear in a bowl game previous year
19	I2	RDWIN%	Road winning percentage
20	I2	SEASTW	Winning percentage for the year
21	I2	TOP25	Winning percentage against AP top 25 teams for the year
22	I3	TSOS	Strength of schedule for the year
23	I3	FR%	Percentage of games played by freshmen class players for the year
24	I3	SO%	Percentage of games played by sophomore class players for the year
25	I3	JR%	Percentage of games played by junior class players for the year
26	I3	SR%	Percentage of games played by senior class players for the year
27	I4	SEASOvUn%	Percentage of times a team went over the O/U[3] in the current season
28	I4	ATSCOV%	Against the spread cover percentage of the team in previous bowl games

(*Continued*)

Application Case 3.4 (Continued)

TABLE 3.5 (Continued)

No	Cat	Variable Name	Description
29	I4	UNDER%	Percentage of times a team went under in previous bowl games
30	I4	OVER%	Percentage of times a team went over in previous bowl games
31	I4	SEASATS%	Percentage of covering against the spread for the current season
32	I5	CONCH	Did the team win their respective conference championship game
33	I5	CONFSOS	Conference strength of schedule
34	I5	CONFWIN%	Conference winning percentage
35	O1	ScoreDiff[4]	Score difference (HomeTeamScore – AwayTeamScore)
36	O2	WinLoss[4]	Whether the home team wins or loses the game

[1]ID: Identifier variables; O1: output variable for regression models; O2: output variable for classification models.
[2]Offense/defense; I2: game outcome; I3: team configuration; I4: against the odds; I5: conference stats.
[3]Over/Under—Whether or not a team will go over or under the expected score difference.
[4]Output variables—ScoreDiff for regression models and WinLoss for binary classification models.

variables 7–34 in Table 3.5), and finally the last two were the output variables (i.e., ScoreDiff—the score difference between the home team and the away team represented with an integer number—and WinLoss—whether the home team won or lost the bowl game represented with a nominal label).

In the formulation of the data set, each row (a.k.a. tuple, case, sample, example, etc.) represented a bowl game, and each column stood for a variable (i.e., identifier/input or output type). To represent the game-related comparative characteristics of the two opponent teams in the input variables, Delen et al. calculated and used the differences between the measures of the home and away teams. All these variable values are calculated from the home team's perspective. For instance, the variable PPG (average number of points a team scored per game) represents the difference between the home team's PPG and away team's PPG. The output variables represent whether the home team wins or loses the bowl game. That is, if the ScoreDiff variable takes a positive integer number, then the home team is expected to win the game by that margin; otherwise (if the ScoreDiff variable takes a negative integer number), the home team is expected to lose the game by that margin. In the case of WinLoss, the value of the output variable is a binary label, "Win" or "Loss," indicating the outcome of the game for the home team.

The Results and Evaluation

In this study, three popular prediction techniques are used to build models (and to compare them to each other): artificial neural networks, decision trees, and support vector machines. These prediction techniques are selected based on their capability of modeling both classification and regression-type prediction problems and their popularity in recently published data mining literature. More details about these popular data mining methods can be found in Chapter 4.

To compare predictive accuracy of all models to one another, the researchers used a stratified *k*-fold cross-validation methodology. In a stratified version of *k*-fold cross-validation, the folds are created in a way that they contain approximately the same proportion of predictor labels (i.e., classes) as the original data set. In this study, the value of *k* is set to 10 (i.e., the complete set of 244 samples are split into 10 subsets, each having about 25 samples), which is a common practice in predictive data mining applications. A graphical depiction of the 10-fold cross-validations was shown earlier in this chapter. To compare the prediction models that were developed using the aforementioned three data mining techniques, the researchers chose to use three common performance criteria: accuracy, sensitivity, and specificity. The simple formulas for these metrics were also explained earlier in this chapter.

TABLE 3.6 Prediction Results for the Direct Classification Methodology

Prediction Method (classification[1])		Confusion Matrix		Accuracy[2] (in %)	Sensitivity (in %)	Specificity (in %)
		Win	Loss			
ANN (MLP)	Win	92	42	75.00	68.66	82.73
	Loss	19	91			
SVM (RBF)	Win	105	29	79.51	78.36	80.91
	Loss	21	89			
DT (C&RT)	Win	113	21	**86.48**	84.33	89.09
	Loss	12	98			

[1]The output variable is a binary categorical variable (Win or Loss).
[2]Differences were significant.

The prediction results of the three modeling techniques are presented in Tables 3.6 and 3.7. Table 3.6 presents the 10-fold cross-validation results of the classification methodology in which the three data mining techniques are formulated to have a binary-nominal output variable (i.e., *WinLoss*). Table 3.7 presents the 10-fold cross-validation results of the regression-based classification methodology in which the three data mining techniques are formulated to have a numerical output variable (i.e., *ScoreDiff*). In the regression-based classification prediction, the numerical output of the models is converted to a classification type by labeling the positive *WinLoss* numbers with a "Win" and negative *WinLoss* numbers with a "Loss" and then tabulating them in the confusion matrixes. Using the confusion matrices, the overall prediction accuracy, sensitivity, and specificity of each model type are calculated and presented in Tables 3.6 and 3.7. As the results indicate, the classification-type prediction methods performed better than regression-based classification-type prediction methodology. Among the three data mining technologies, classification and regression trees produced better prediction accuracy in both prediction methodologies. Overall, classification and regression tree classification models produced a 10-fold cross-validation accuracy of 86.48 percent followed by support vector machines

TABLE 3.7 Prediction Results for the Regression-Based Classification Methodology

Prediction Method (regression based[1])		Confusion Matrix		Accuracy[2]	Sensitivity	Specificity
		Win	Loss			
ANN (MLP)	Win	94	40	72.54	70.15	75.45
	Loss	27	83			
SVM (RBF)	Win	100	34	74.59	74.63	74.55
	Loss	28	82			
DT (C&RT)	Win	106	28	77.87	76.36	79.10
	Loss	26	84			

[1]The output variable is a numerical/integer variable (point-diff).
[2]Differences were sig $p < 0.01$.

(*Continued*)

Application Case 3.4 (Continued)

(with a 10-fold cross-validation accuracy of 79.51 percent) and neural networks (with a 10-fold cross-validation accuracy of 75.00 percent). Using a *t*-test, researchers found that these accuracy values were significantly different at 0.05 alpha level; that is, the decision tree is a significantly better predictor of this domain than the neural network and support vector machine, and the support vector machine is a significantly better predictor than neural networks.

The results of the study showed that the classification-type models predict the game outcomes better than regression-based classification models. Even though these results are specific to the application domain and the data used in this study and therefore should not be generalized beyond the scope of the study, they are exciting because decision trees are not only the best predictors but also the best in understanding and deployment, compared to the other two machine-learning techniques employed in this study. More details about this study can be found in Delen et al. (2012).

Questions for Case 3.4

1. What are the foreseeable challenges in predicting sporting event outcomes (e.g., college bowl games)?
2. How did the researchers formulate/design the prediction problem (i.e., what were the inputs and output, and what was the representation of a single sample—row of data)?
3. How successful were the prediction results? What else can they do to improve the accuracy?

Sources: D. Delen, D. Cogdell, and N. Kasap, "A Comparative Analysis of Data Mining Methods in Predicting NCAA Bowl Outcomes," *International Journal of Forecasting, 28*, 2012, pp. 543–552; K. M. Freeman, and R. M. Brewer, "The Politics of American College Football," *Journal of Applied Business and Economics, 18*(2), 2016, pp. 97–101.

daily sales totals for a grocery store. Often, time series are visualized using a line chart. Figure 3.17 shows an example time series of sales volumes for the years 2008 through 2012 on a quarterly basis.

Time-series forecasting is the use of mathematical modeling to predict future values of the variable of interest based on previously observed values. The time-series plots/charts look and feel very similar to simple linear regression in that, as was the case in simple linear regression, in time series there are two variables: the response variable and the time variable presented in a scatter plot. Beyond this appearance similarity, there is hardly any other commonality between the two. Although regression analysis is often employed in testing theories to see if current values of one or more explanatory variables explain (and hence predict) the response variable, the time-series models are focused on extrapolating on their time-varying behavior to estimate the future values.

Time-series-forecasting assumes that all of the explanatory variables are aggregated into the response variable as a time-variant behavior. Therefore, capturing the time-variant behavior is the way to predict the future values of the response variable. To do that, the pattern is analyzed and decomposed into its main components: random variations, time trends, and seasonal cycles. The time-series example shown in Figure 3.17 illustrates all of these distinct patterns.

The techniques used to develop time-series forecasts range from very simple (the naïve forecast that suggests today's forecast is the same as yesterday's actual) to very complex like ARIMA (a method that combines autoregressive and moving average patterns in data). Most popular techniques are perhaps the averaging methods that include simple average, moving average, weighted moving average, and exponential smoothing. Many of these techniques also have advanced versions when seasonality and trend can also be taken into account for better and more accurate forecasting. The accuracy of a method is usually assessed by computing its error (calculated deviation between actuals and forecasts for the past observations) via mean absolute error (MAE), mean squared error (MSE), or mean absolute percent error (MAPE). Even though they all use the same

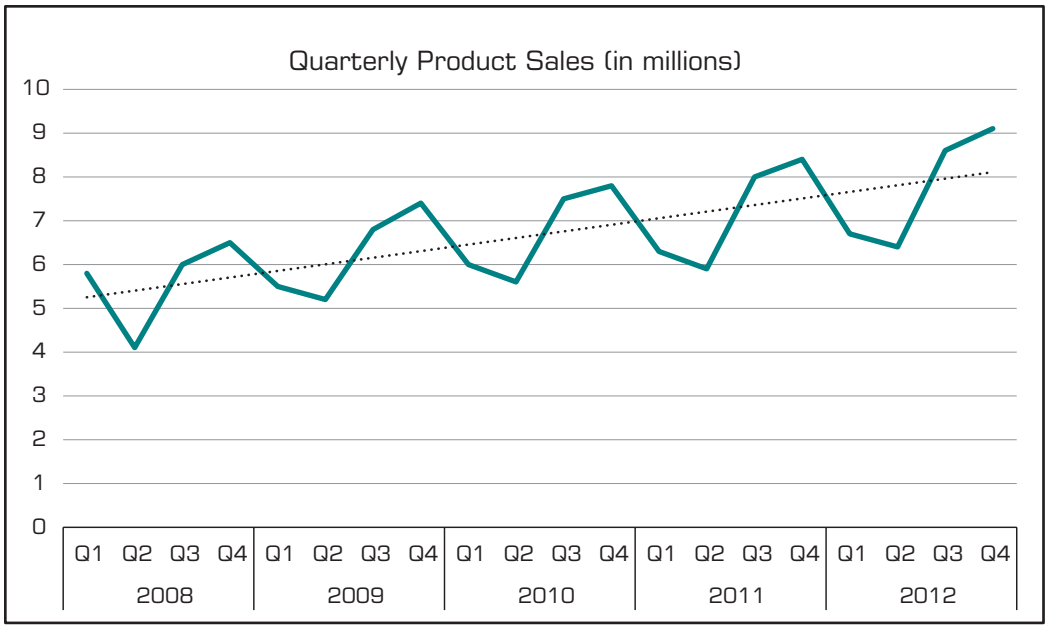


FIGURE 3.17 **A Sample Time Series of Data on Quarterly Sales Volumes.**

core error measure, these three assessment methods emphasize different aspects of the error, some penalizing larger errors more so than the others.

SECTION 3.6 REVIEW QUESTIONS

1. What is regression, and what statistical purpose does it serve?
2. What are the commonalities and differences between regression and correlation?
3. What is OLS? How does OLS determine the linear regression line?
4. List and describe the main steps to follow in developing a linear repression model.
5. What are the most commonly pronounced assumptions for linear regression?
6. What is logistics regression? How does it differ from linear regression?
7. What is time series? What are the main forecasting techniques for time-series data?

3.7 BUSINESS REPORTING

Decision makers need information to make accurate and timely decisions. Information is essentially the contextualization of data. In addition to statistical means that were explained in the previous section, information (descriptive analytics) can also be obtained using OLTP systems (see the simple taxonomy of descriptive analytics in Figure 3.7). The information is usually provided to decision makers in the form of a written report (digital or on paper), although it can also be provided orally. Simply put, a **report** is any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it whenever and wherever. It is typically a document that contains information (usually driven from data) organized in a narrative, graphic, and/or tabular form, prepared periodically (recurring) or on an as-needed (ad hoc) basis, referring to specific time periods, events, occurrences, or subjects. Business reports can fulfill many different (but often related) functions. Here are a few of the most prevailing ones:

- To ensure that all departments are functioning properly.
- To provide information.

- To provide the results of an analysis.
- To persuade others to act.
- To create an organizational memory (as part of a knowledge management system).

Business reporting (also called OLAP or BI) is an essential part of the larger drive toward improved, evidence-based, optimal managerial decision making. The foundation of these **business reports** is various sources of data coming from both inside and outside the organization (OLTP systems). Creation of these reports involves extract, transform, and load (ETL) procedures in coordination with a data warehouse and then using one or more reporting tools.

Due to the rapid expansion of IT coupled with the need for improved competitiveness in business, there has been an increase in the use of computing power to produce unified reports that join different views of the enterprise in one place. Usually, this reporting process involves querying structured data sources, most of which were created using different logical data models and data dictionaries, to produce a human-readable, easily digestible report. These types of business reports allow managers and coworkers to stay informed and involved, review options and alternatives, and make informed decisions. Figure 3.18 shows the continuous cycle of data acquisition → information generation → decision-making → business process management. Perhaps the most critical task in this cyclical process is the reporting (i.e., information generation)—converting data from different sources into actionable information.

Key to any successful report are clarity, brevity, completeness, and correctness. The nature of the report and the level of importance of these success factors changes significantly based on for whom the report is created. Most of the research in effective reporting is dedicated to internal reports that inform stakeholders and decision makers within the organization. There are also external reports between businesses and the government (e.g., for tax purposes or for regular filings to the Securities and Exchange Commission). Even though there is a wide variety of business reports, the ones that are often used for managerial purposes can be grouped into three major categories (Hill, 2016).

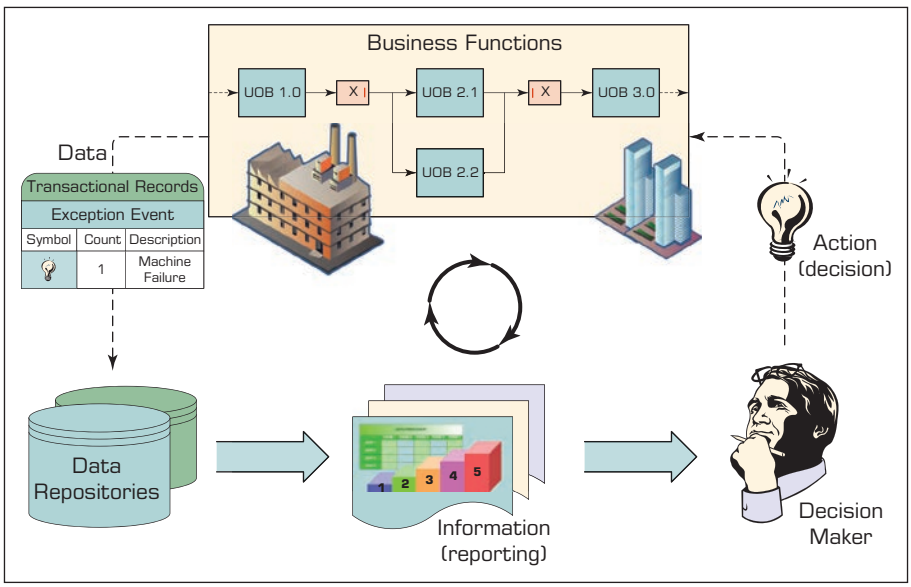


FIGURE 3.18 The Role of Information Reporting in Managerial Decision Making.

METRIC MANAGEMENT REPORTS In many organizations, business performance is managed through outcome-oriented metrics. For external groups, these are service-level agreements. For internal management, they are **key performance indicators (KPIs)**. Typically, there are enterprise-wide agreed upon targets to be tracked against over a period of time. They can be used as part of other management strategies such as Six Sigma or total quality management.

DASHBOARD-TYPE REPORTS A popular idea in business reporting in recent years has been to present a range of different performance indicators on one page like a dashboard in a car. Typically, dashboard vendors would provide a set of predefined reports with static elements and fixed structure but also allow for customization of the dashboard widgets, views, and set targets for various metrics. It is common to have color-coded traffic lights defined for performance (red, orange, green) to draw management's attention to particular areas. A more detailed description of dashboards can be found in a later section of this chapter.

BALANCED SCORECARD–TYPE REPORTS This is a method developed by Kaplan and Norton that attempts to present an integrated view of success in an organization. In addition to financial performance, balanced scorecard–type reports also include customer, business process, and learning and growth perspectives. More details on balanced scorecards are provided in a later section in this chapter.

Application Case 3.5 is an example to illustrate the power and the utility of automated report generation for a large (and, at a time of natural crisis, somewhat chaotic) organization such as the Federal Emergency Management Agency.

Application Case 3.5 Flood of Paper Ends at FEMA

Staff at the Federal Emergency Management Agency (FEMA), the U.S. federal agency that coordinates disaster response when the president declares a national disaster, always got two floods at once. First, water covered the land. Next, a flood of paper required to administer the National Flood Insurance Program (NFIP) covered their desks—pallets and pallets of green-striped reports poured off a mainframe printer and into their offices. Individual reports were sometimes 18 inches thick with a nugget of information about insurance claims, premiums, or payments buried in them somewhere.

Bill Barton and Mike Miles do not claim to be able to do anything about the weather, but the project manager and computer scientist, respectively, from Computer Sciences Corporation (CSC) have used WebFOCUS software from Information Builders to turn back the flood of paper generated by the NFIP. The program allows the government to work with national insurance companies to collect flood insurance premiums and pay claims for flooding in communities that adopt flood control measures. As a result of CSC's work, FEMA staffs no longer leaf through paper reports to find the data they need. Instead, they browse insurance data posted on NFIP's BureauNet intranet site, select just the information they want to see, and get an on-screen report or download the data as a spreadsheet. And that is only the start of the savings that WebFOCUS has provided. The number of times that NFIP staff ask CSC for special reports has dropped in half because NFIP staff can generate many of the special reports they need without calling on a programmer to develop them. Then there is the cost of creating BureauNet in the first place. Barton estimates that using conventional Web and database software to export data from FEMA's mainframe,

(*Continued*)

Application Case 3.5 (Continued)

store it in a new database, and link that to a Web server would have cost about 100 times as much—more than $500,000—and taken about two years to complete compared with the few months Miles spent on the WebFOCUS solution.

When Tropical Storm Allison, a huge slug of sodden, swirling cloud, moved out of the Gulf of Mexico onto the Texas and Louisiana coastline in June 2001, it killed 34 people, most from drowning; damaged or destroyed 16,000 homes and businesses; and displaced more than 10,000 families. President George W. Bush declared 28 Texas counties disaster areas, and FEMA moved in to help. This was the first serious test for BureauNet, and it delivered. This first comprehensive use of BureauNet resulted in FEMA field staff readily accessing what they needed when they needed it and asking for many new types of reports. Fortunately, Miles and WebFOCUS were up to the task. In some cases, Barton says, "FEMA would ask for a new type of report one day, and Miles would have it on BureauNet the next day, thanks to the speed with which he could create new reports in WebFOCUS."

The sudden demand on the system had little impact on its performance, noted Barton. "It handled the demand just fine," he says. "We had no problems with it at all. And it made a huge difference to FEMA and the job they had to do. They had never had that level of access before, never had been able to just click on their desktop and generate such detailed and specific reports."

Questions for Case 3.5

1. What is FEMA, and what does it do?
2. What are the main challenges that FEMA faces?
3. How did FEMA improve its inefficient reporting practices?

Source: Used with permission from Information Builders. Useful information flows at disaster response agency. **informationbuilders.com/applications/fema** (accessed July 2018); and **fema.gov**.

SECTION 3.7 REVIEW QUESTIONS

1. What is a report? What are reports used for?
2. What is a business report? What are the main characteristics of a good business report?
3. Describe the cyclic process of management, and comment on the role of business reports.
4. List and describe the three major categories of business reports.
5. What are the main components of a business reporting system?

3.8 DATA VISUALIZATION

Data visualization (or more appropriately, information visualization) has been defined as "the use of visual representations to explore, make sense of, and communicate data" (Few, 2007). Although the name that is commonly used is *data visualization*, usually what this means is information visualization. Because information is the aggregation, summarization, and contextualization of data (raw facts), what is portrayed in visualizations is the information, not the data. However, because the two terms *data visualization* and *information visualization* are used interchangeably and synonymously, in this chapter we will follow suit.

Data visualization is closely related to the fields of information graphics, information visualization, scientific visualization, and statistical graphics. Until recently, the major forms of data visualization available in both BI applications have included charts and graphs as well as the other types of visual elements used to create scorecards and dashboards.

To better understand the current and future trends in the field of data visualization, it helps to begin with some historical context.

Brief History of Data Visualization

Despite the fact that predecessors to data visualization date back to the second century AD, most developments have occurred in the last two and a half centuries, predominantly during the last 30 years (Few, 2007). Although visualization has not been widely recognized as a discipline until fairly recently, today's most popular visual forms date back a few centuries. Geographical exploration, mathematics, and popularized history spurred the creation of early maps, graphs, and timelines as far back as the 1600s, but William Playfair is widely credited as the inventor of the modern chart, having created the first widely distributed line and bar charts in his *Commercial and Political Atlas of 1786* and what is generally considered to be the first time-series line chart in his *Statistical Breviary* published in 1801 (see Figure 3.19).

Perhaps the most notable innovator of information graphics during this period was Charles Joseph Minard, who graphically portrayed the losses suffered by Napoleon's army in the Russian campaign of 1812 (see Figure 3.20). Beginning at the Polish–Russian border, the thick band shows the size of the army at each position. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. Popular visualization expert, author, and critic Edward Tufte says that this "may well be the best statistical graphic ever drawn." In this graphic, Minard managed to simultaneously represent several data dimensions (the size of the army, direction of movement, geographic locations, outside temperature, etc.) in an artistic and informative

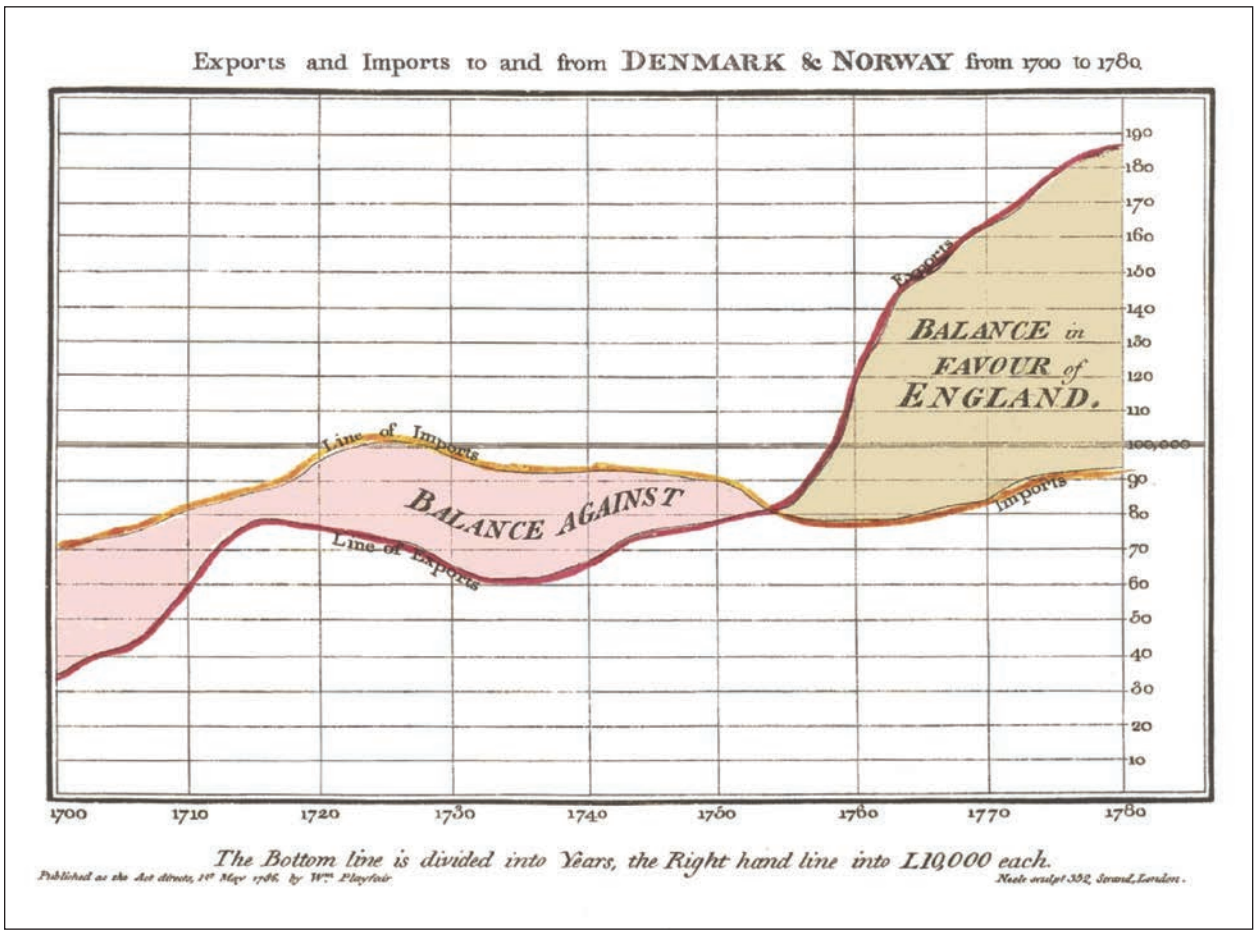


FIGURE 3.19 **The First Time-Series Line Chart Created by William PlayFair in 1801.**

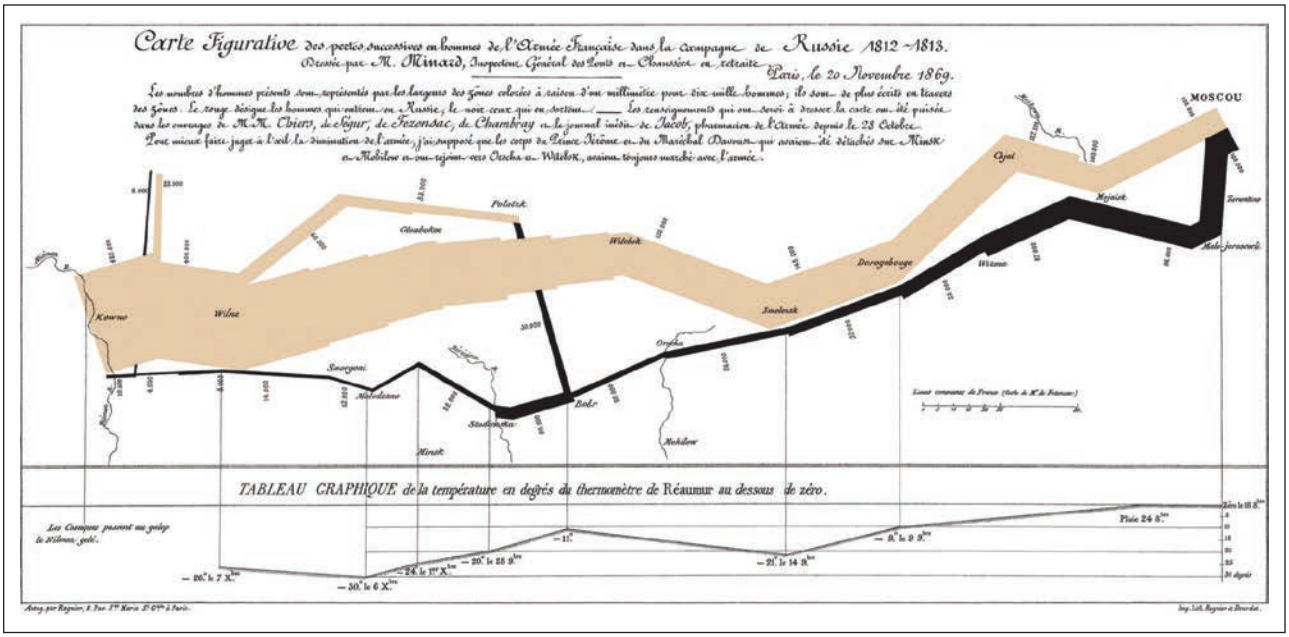


FIGURE 3.20 **Decimation of Napoleon's Army during the 1812 Russian Campaign.**

manner. Many more excellent visualizations were created in the 1800s, and most of them are chronicled on Tufte's Web site (**edwardtufte.com**) and his visualization books.

The 1900s saw the rise of a more formal, empirical attitude toward visualization, which tended to focus on aspects such as color, value scales, and labeling. In the mid-1900s, cartographer and theorist Jacques Bertin published his *Semiologie Graphique*, which some say serves as the theoretical foundation of modern information visualization. Although most of his patterns are either outdated by more recent research or completely inapplicable to digital media, many are still very relevant.

In the 2000s, the Internet emerged as a new medium for visualization and brought with it many new tricks and capabilities. Not only has the worldwide, digital distribution of both data and visualization made them more accessible to a broader audience (raising visual literacy along the way), but also it has spurred the design of new forms that incorporate interaction, animation, and graphics-rendering technology unique to screen media and real-time data feeds to create immersive environments for communicating and consuming data.

Companies and individuals are, seemingly all of a sudden, interested in data; that interest has in turn sparked a need for visual tools that help them understand it. Cheap hardware sensors and do-it-yourself frameworks for building your own system are driving down the costs of collecting and processing data. Countless other applications, software tools, and low-level code libraries are springing up to help people collect, organize, manipulate, visualize, and understand data from practically any source. The Internet has also served as a fantastic distribution channel for visualizations; a diverse community of designers, programmers, cartographers, tinkerers, and data wonks has assembled to disseminate all sorts of new ideas and tools for working with data in both visual and nonvisual forms.

Google Maps has also single-handedly democratized both the interface conventions (click to pan, double-click to zoom) and the technology (256-pixel square map tiles with predictable file names) for displaying interactive geography online to the extent that most people just know what to do when they are presented with a map online. Flash has served well as a cross-browser platform on which to design and develop rich, beautiful Internet applications incorporating interactive data visualization and maps; now, new

browser-native technologies such as canvas and SVG (sometimes collectively included under the umbrella of HTML5) are emerging to challenge Flash's supremacy and extend the reach of dynamic visualization interfaces to mobile devices.

The future of data/information visualization is very hard to predict. We can only extrapolate from what has already been invented: more three-dimensional visualization, immersive experience with multidimensional data in a virtual reality environment, and holographic visualization of information. There is a pretty good chance that we will see something that we have never seen in the information visualization realm invented before the end of this decade. Application Case 3.6 shows how visual analytics/reporting tools such as Tableau can help facilitate effective and efficient decision making through information/insight creation and sharing.

Application Case 3.6

Macfarlan Smith Improves Operational Performance Insight with Tableau Online

The Background

Macfarlan Smith has earned its place in medical history. The company held a royal appointment to provide medicine to Her Majesty Queen Victoria and supplied groundbreaking obstetrician Sir James Simpson with chloroform for his experiments in pain relief during labor and delivery. Today, Macfarlan Smith is a subsidiary of the Fine Chemical and Catalysts division of Johnson Matthey plc. The pharmaceutical manufacturer is the world's leading manufacturer of opiate narcotics such as codeine and morphine.

Every day, Macfarlan Smith is making decisions based on its data. The company collects and analyzes manufacturing operational data, for example, to allow it to meet continuous improvement goals. Sales, marketing, and finance rely on data to identify new pharmaceutical business opportunities, grow revenues, and satisfy customer needs. Additionally, the company's manufacturing facility in Edinburgh needs to monitor, trend, and report quality data to ensure the identity, quality, and purity of its pharmaceutical ingredients for customers and regulatory authorities such as the U.S. FDA and others as part of current good manufacturing practice (CGMP).

Challenges: Multiple Sources of Truth and Slow, Onerous Reporting Processes

The process of gathering that data, making decisions, and reporting was not easy, though. The data were scattered across the business including in the company's bespoke enterprise resource planning (ERP) platform, inside legacy departmental databases such as SQL, Access databases, and stand-alone spreadsheets. When those data were needed for decision making, excessive time and resources were devoted to extracting the data, integrating them, and presenting them in a spreadsheet or other presentation outlet.

Data quality was another concern. Because teams relied on their own individual sources of data, there were multiple versions of the truth and conflicts between the data. And it was sometimes hard to tell which version of the data was correct and which was not.

It didn't stop there. Even once the data had been gathered and presented, making changes "on the fly" was slow and difficult. In fact, whenever a member of the Macfarlan Smith team wanted to perform trend or other analysis, the changes to the data needed to be approved. The end result was that the data were frequently out of date by the time they were used for decision making.

Liam Mills, Head of Continuous Improvement at Macfarlan Smith highlights a typical reporting scenario:

> One of our main reporting processes is the "Corrective Action and Preventive Action," or CAPA, which is an analysis of Macfarlan Smith's manufacturing processes taken to eliminate causes of non-conformities or other undesirable situations. Hundreds of hours every month were devoted to pulling data together for CAPA—and it took days to produce each

(*Continued*)

Application Case 3.6 (Continued)

report. Trend analysis was tricky too, because the data was static. In other reporting scenarios, we often had to wait for spreadsheet pivot table analysis; which was then presented on a graph, printed out, and pinned to a wall for everyone to review.

Slow, labor-intensive reporting processes, different versions of the truth, and static data were all catalysts for change. "Many people were frustrated because they believed they didn't have a complete picture of the business," says Mills. "We were having more and more discussions about issues we faced—when we should have been talking about business intelligence reporting."

The Solution: Interactive Data Visualizations

One of the Macfarlan Smith team had previous experience in using Tableau and recommended Mills explore the solution further. A free trial of Tableau Online quickly convinced Mills that the hosted interactive data visualization solution could conquer the data battles the company was facing.

"I was won over almost immediately," he says. "The ease of use, the functionality and the breadth of data visualizations are all very impressive. And of course being a software-as-a-service (SaaS)-based solution, there's no technology infrastructure investment, we can be live almost immediately, and we have the flexibility to add users whenever we need."

One of the key questions that needed to be answered concerned the security of the online data. "Our parent company Johnson Matthey has a cloud-first strategy, but has to be certain that any hosted solution is completely secure. Tableau Online features like single sign-on and allowing only authorized users to interact with the data provide that watertight security and confidence."

The other security question that Macfarlan Smith and Johnson Matthey wanted answered was this: Where are the data physically stored? Mills again: "We are satisfied Tableau Online meets our criteria for data security and privacy. The data and workbooks are all hosted in Tableau's new Dublin data center, so it never leaves Europe."

Following a six-week trial, the Tableau sales manager worked with Mills and his team to build a business case for Tableau Online. The management team approved it almost straight away, and a pilot program involving 10 users began. The pilot involved a manufacturing quality improvement initiative: looking at deviations from the norm, such as when a heating device used in the opiate narcotics manufacturing process exceeds a temperature threshold. From this, a "quality operations" dashboard was created to track and measure deviations and put in place measures to improve operational quality and performance.

"That dashboard immediately signaled where deviations might be. We weren't ploughing through rows of data—we reached answers straight away," says Mills.

Throughout this initial trial and pilot, the team used Tableau training aids, such as the free training videos, product walk-throughs, and live online training. They also participated in a two-day "fundamentals training" event in London. According to Mills, "The training was expert, precise and pitched just at the right level. It demonstrated to everyone just how intuitive Tableau Online is. We can visualize 10 years' worth of data in just a few clicks." The company now has five Tableau Desktop users and up to 200 Tableau Online licensed users.

Mills and his team particularly like the Tableau Union feature in Version 9.3, which allows them to piece together data that have been split into little files. "It's sometimes hard to bring together the data we use for analysis. The Union feature lets us work with data spread across multiple tabs or files, reducing the time we spend on prepping the data," he says.

The Results: Cloud Analytics Transform Decision Making and Reporting

By standardizing on Tableau Online, Macfarlan Smith has transformed the speed and accuracy of its decision making and business reporting. This includes:

- New interactive dashboards can be produced within one hour. Previously, it used to take days to integrate and present data in a static spreadsheet.
- The CAPA manufacturing process report, which used to absorb hundreds of man-hours every month and days to produce, can now be produced in minutes—with insights shared in the cloud.

- Reports can be changed and interrogated "on the fly" quickly and easily, without technical intervention. Macfarlan Smith has the flexibility to publish dashboards with Tableau Desktop and share them with colleagues, partners, or customers.
- The company has one, single, trusted version of the truth.
- Macfarlan Smith is now having discussions about its data—not about the issues surrounding data integration and data quality.
- New users can be brought online almost instantly—and there's no technical infrastructure to manage.

Following this initial success, Macfarlan Smith is now extending Tableau Online to financial reporting, supply chain analytics, and sales forecasting. Mills concludes, "Our business strategy is now based on data-driven decisions, not opinions. The interactive visualizations enable us to spot trends instantly, identify process improvements and take business intelligence to the next level. I'll define my career by Tableau."

Questions for Case 3.6

1. What were the data and reporting related challenges that Macfarlan Smith faced?
2. What were the solution and the obtained results/benefits?

Source: Tableau Customer Case Study, "Macfarlan Smith improves operational performance insight with Tableau Online," **http://www.tableau.com/stories/customer/macfarlan-smith-improves-operational-performance-insight-tableau-online** (accessed June 2018). Used with permission from Tableau Software, Inc.

SECTION 3.8 REVIEW QUESTIONS

1. What is data visualization? Why is it needed?
2. What are the historical roots of data visualization?
3. Carefully analyze Charles Joseph Minard's graphical portrayal of Napoleon's march. Identify and comment on all the information dimensions captured in this ancient diagram.
4. Who is Edward Tufte? Why do you think we should know about his work?
5. What do you think is the "next big thing" in data visualization?

3.9 DIFFERENT TYPES OF CHARTS AND GRAPHS

Often end users of business analytics systems are not sure what type of chart or graph to use for a specific purpose. Some charts or graphs are better at answering certain types of questions. Some look better than others. Some are simple; some are rather complex and crowded. What follows is a short description of the types of charts and/or graphs commonly found in most business analytics tools and the types of questions they are better at answering/analyzing. This material is compiled from several published articles and other literature (Abela, 2008; Hardin et al., 2012; SAS, 2014).

Basic Charts and Graphs

What follows are the basic charts and graphs that are commonly used for information visualization.

LINE CHART The line chart is perhaps the most frequently used graphical visuals for time-series data. Line charts (or line graphs) show the relationship between two variables; they are most often used to track changes or trends over time (having one of the variables set to time on the *x-axis*). Line charts sequentially connect individual data points to help infer changing trends over a period of time. Line charts are often used to show time-dependent changes in the values of some measure, such as changes in a specific

stock price over a five-year period or changes in the number of daily customer service calls over a month.

BAR CHART The bar chart is among the most basic visuals used for data representation. They are effective when you have nominal data or numerical data that split nicely into different categories so you can quickly see comparative results and trends within your data. Bar charts are often used to compare data across multiple categories such as the percentage of advertising spending by departments or by product categories. Bar charts can be vertically or horizontally oriented. They can also be stacked on top of each other to show multiple dimensions in a single chart.

PIE CHART The **pie chart** is visually appealing, as the name implies, pie-looking charts. Because they are so visually attractive, they are often incorrectly used. Pie charts should be used only to illustrate relative proportions of a specific measure. For instance, they can be used to show the relative percentage of an advertising budget spent on different product lines, or they can show relative proportions of majors declared by college students in their sophomore year. If the number of categories to show is more than just a few (say more than four), one should seriously consider using a bar chart instead of a pie chart.

SCATTER PLOT The **scatter plot** is often used to explore the relationship between two or three variables (in 2D or 3D visuals). Because scatter plots are visual exploration tools, translating more than three variables into more than three dimensions is not easily achievable. Scatter plots are an effective way to explore the existence of trends, concentrations, and outliers. For instance, in a two-variable (two-axis) graph, a scatter plot can be used to illustrate the co-relationship between age and weight of heart disease patients, or it can illustrate the relationship between the number of customer care representatives and the number of open customer service claims. Often, a trend line is superimposed on a two-dimensional scatter plot to illustrate the nature of the relationship.

BUBBLE CHART The **bubble chart** is often an enhanced version of scatter plots. Bubble charts, though, are not a new visualization type; instead, they should be viewed as a technique to enrich data illustrated in scatter plots (or even geographic maps). By varying the size and/or color of the circles, one can add additional data dimensions, offering more enriched meaning about the data. For instance, a bubble chart can be used to show a competitive view of college-level class attendance by major and by time of the day, and it can be used to show profit margin by product type and by geographic region.

Specialized Charts and Graphs

The graphs and charts that we review in this section are either derived from the basic charts as special cases or they are relatively new and are specific to a problem type and/or an application area.

HISTOGRAM Graphically speaking, a **histogram** looks just like a bar chart. The difference between histograms and generic bar charts is the information that is portrayed. Histograms are used to show the frequency distribution of one variable or several variables. In a histogram, the *x-axis* is often used to show the categories or ranges, and the *y-axis* is used to show the measures/values/frequencies. Histograms show the distributional shape of the data. That way, one can visually examine whether the data are normally or exponentially distributed. For instance, one can use a histogram to illustrate the

exam performance of a class, to show distribution of the grades as well as comparative analysis of individual results, or to show the age distribution of the customer base.

GANTT CHART A Gantt chart is a special case of horizontal bar charts used to portray project timelines, project tasks/activity durations, and overlap among the tasks/activities. By showing start and end dates/times of tasks/activities and the overlapping relationships, Gantt charts provide an invaluable aid for management and control of projects. For instance, Gantt charts are often used to show project timelines, task overlaps, relative task completions (a partial bar illustrating the completion percentage inside a bar that shows the actual task duration), resources assigned to each task, milestones, and deliverables.

PERT CHART The PERT chart (also called a *network diagram*) is developed primarily to simplify the planning and scheduling of large and complex projects. A PERT chart shows precedence relationships among project activities/tasks. It is composed of nodes (represented as circles or rectangles) and edges (represented with directed arrows). Based on the selected PERT chart convention, either nodes or the edges can be used to represent the project activities/tasks (activity-on-node versus activity-on-arrow representation schema).

GEOGRAPHIC MAP When the data set includes any kind of location data (e.g., physical addresses, postal codes, state names or abbreviations, country names, latitude/longitude, or some type of custom geographic encoding), it is better and more informative to see the data on a map. Maps usually are used in conjunction with other charts and graphs rather than by themselves. For instance, one can use maps to show the distribution of customer service requests by product type (depicted in pie charts) by geographic locations. Often a large variety of information (e.g., age distribution, income distribution, education, economic growth, population changes) can be portrayed in a geographic map to help decide where to open a new restaurant or a new service station. These types of systems are often called *geographic information systems* (GIS).

BULLET A bullet graph is often used to show progress toward a goal. This graph is essentially a variation of a bar chart. Often bullet graphs are used in place of gauges, meters, and thermometers in a dashboard to more intuitively convey the meaning within a much smaller space. Bullet graphs compare a primary measure (e.g., year-to-date revenue) to one or more other measures (e.g., annual revenue target) and present this in the context of defined performance metrics (e.g., sales quotas). A bullet graph can intuitively illustrate how the primary measure is performing against overall goals (e.g., how close a sales representative is to achieving his or her annual quota).

HEAT MAP The heat map is a great visual to illustrate the comparison of continuous values across two categories using color. The goal is to help the user quickly see where the intersection of the categories is strongest and weakest in terms of numerical values of the measure being analyzed. For instance, one can use a heat map to show segmentation analysis of target markets where the measure (color gradient would be the purchase amount) and the dimensions would be age and income distribution.

HIGHLIGHT TABLE The highlight table is intended to take heat maps one step further. In addition to showing how data intersect by using color, highlight tables add a number on top to provide additional detail. That is, they are two-dimensional tables with cells populated with numerical values and gradients of colors. For instance, one can show sales representatives' performance by product type and by sales volume.

TREE MAP A tree map displays hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing subbranches. A leaf node's rectangle has an area proportional to a specified dimension on the data. Often the leaf nodes are colored to show a separate dimension of the data. When the color and size dimensions are correlated in some way with the tree structure, one can often easily see patterns that would be difficult to spot in other ways, such as a certain color that is particularly relevant. A second advantage of tree maps is that, by construction, they make efficient use of space. As a result, they can legibly display thousands of items on the screen simultaneously.

Which Chart or Graph Should You Use?

Which chart or graph that we explained in the previous section is the best? The answer is rather easy: There is not one best chart or graph because if there were, we would not have so many chart and graph types. They all have somewhat different data representation "skills." Therefore, the right question should be, "Which chart or graph is the best for a given task?" The capabilities of the charts given in the previous section can help in selecting and using the proper chart/graph for a specific task, but doing so still is not easy to sort out. Several different chart/graph types can be used for the same visualization task. One rule of thumb is to select and use the simplest one from the alternatives to make it easy for the intended audience to understand and digest.

Although there is not a widely accepted, all-encompassing chart selection algorithm or chart/graph taxonomy, Figure 3.21 presents a rather comprehensive and highly logical

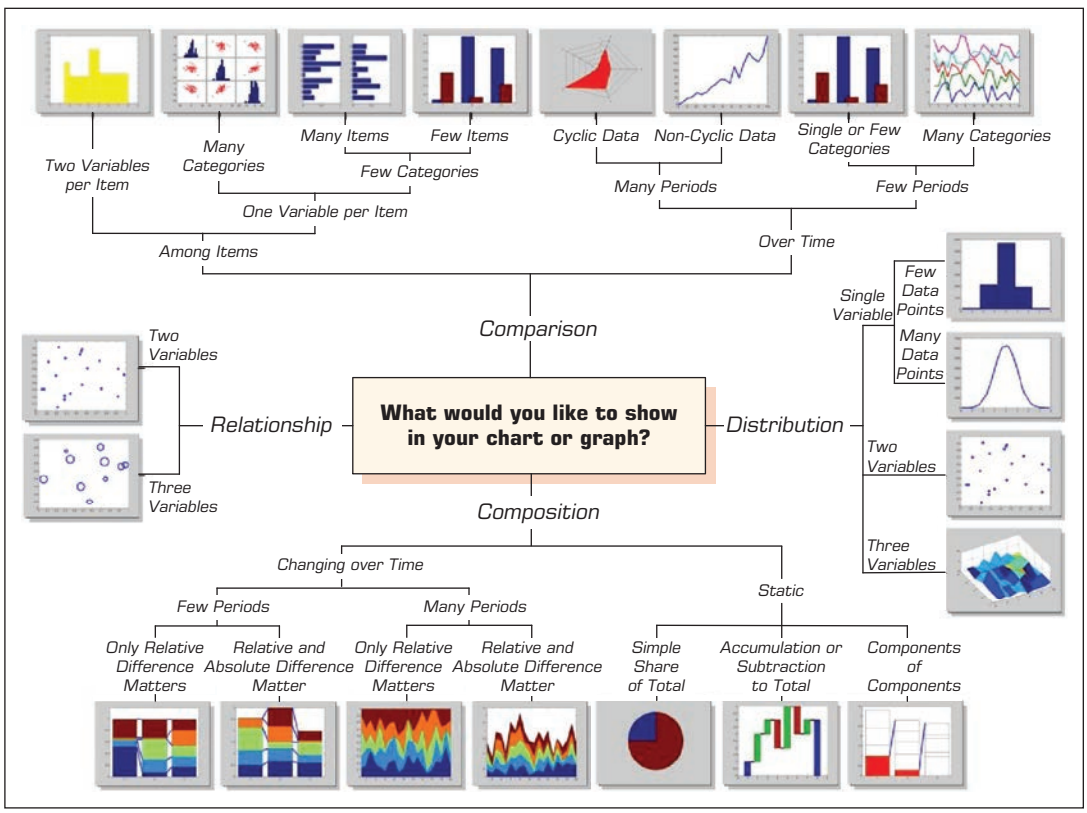


FIGURE 3.21 A Taxonomy of Charts and Graphs. *Source:* Adapted from Abela, A. (2008). *Advanced Presentations by Design: Creating Communication That Drives Action*. New York: Wiley.

organization of chart/graph types in a taxonomy-like structure (the original version was published in Abela, 2008). The taxonomic structure is organized around the questions of "What would you like to show in your chart or graph?"—that is, what the purpose of the chart or graph will be. At that level, the taxonomy divides the purpose into four different types—relationship, comparison, distribution, and composition—and further divides the branches into subcategories based on the number of variables involved and time dependency of the visualization.

Even though these charts and graphs cover a major part of what is commonly used in information visualization, they by no means cover all. Today, one can find many other specialized graphs and charts that serve a specific purpose. Furthermore, the current trend is to combine/hybridize and animate these charts for better-looking and more intuitive visualization of today's complex and volatile data sources. For instance, the interactive, animated, bubble charts available at the Gapminder Web site (**gapminder.org**) provide an intriguing way of exploring world health, wealth, and population data from a multidimensional perspective. Figure 3.22 depicts the types of displays available at that site. In this graph, population size, life expectancy, and per capita income at the continent level are shown; also given is a time-varying animation that shows how these variables change over time.

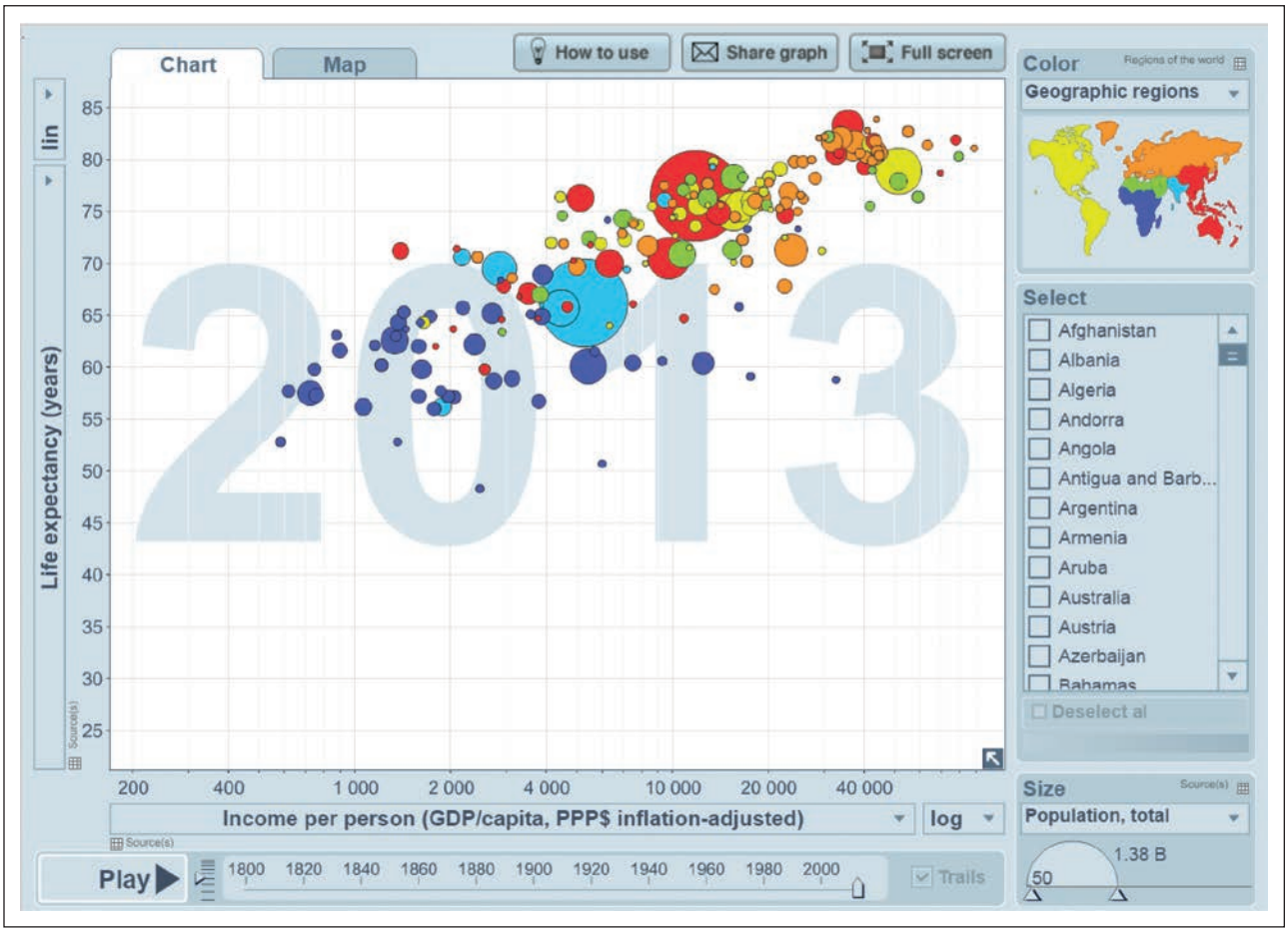


FIGURE 3.22 **A Gapminder Chart That Shows the Wealth and Health of Nations.** *Source:* gapminder.org.

SECTION 3.9 REVIEW QUESTIONS

1. Why do you think there are many different types of charts and graphs?
2. What are the main differences among line, bar, and pie charts? When should you use one over the others?
3. Why would you use a geographic map? What other types of charts can be combined with it?
4. Find and explain the role of two types of charts that are not covered in this section.

3.10 EMERGENCE OF VISUAL ANALYTICS

As Seth Grimes (2009a, b) has noted, there is a "growing palate" of data visualization techniques and tools that enable the users of business analytics and BI systems to better "communicate relationships, add historical context, uncover hidden correlations, and tell persuasive stories that clarify and call to action." The latest Magic Quadrant for Business Intelligence and Analytics Platforms released by Gartner in February 2016 further emphasizes the importance of data visualization in BI and analytics. As the chart in Figure 3.23 shows, all solution

FIGURE 3.23 Magic Quadrant for Business Intelligence and Analytics Platforms. *Source:* Used with permission from Gartner Inc.

providers in the *Leaders* and *Visionaries* quadrants are either relatively recently founded information visualization companies (e.g., Tableau Software, QlikTech) or well-established large analytics companies (e.g., Microsoft, SAS, IBM, SAP, MicroStrategy, Alteryx) that are increasingly focusing their efforts on information visualization and visual analytics. More details on Gartner's latest Magic Quadrant are given in Technology Insights 3.2.

In BI and analytics, the key challenges for visualization have revolved around the intuitive representation of large, complex data sets with multiple dimensions and measures. For the most part, the typical charts, graphs, and other visual elements used in these applications usually involve two dimensions, sometimes three, and fairly small subsets of data sets. In contrast, the data in these systems reside in a data warehouse. At a

TECHNOLOGY INSIGHTS 3.2 Gartner Magic Quadrant for Business Intelligence and Analytics Platforms

Gartner, Inc., the creator of Magic Quadrants, is the leading IT research and advisory company publically traded in the United States with over $2 billion annual revenues in 2015. Founded in 1979, Gartner has 7,600 associates, including 1,600 research analysts and consultants and numerous clients in 90 countries.

Magic Quadrant is a research method designed and implemented by Gartner to monitor and evaluate the progress and positions of companies in a specific, technology-based market. By applying a graphical treatment and a uniform set of evaluation criteria, Magic Quadrant helps users to understand how technology providers are positioned within a market.

Gartner changed the name of this Magic Quadrant from "Business Intelligence Platforms" to "Business Intelligence and Analytics Platforms" to emphasize the growing importance of analytics capabilities to the information systems that organizations are now building. Gartner defines the BI and analytics platform market as a software platform that delivers 15 capabilities across three categories: integration, information delivery, and analysis. These capabilities enable organizations to build precise systems of classification and measurement to support decision making and improve performance.

Figure 3.23 illustrates the latest Magic Quadrant for Business Intelligence and Analytics Platforms. Magic Quadrant places providers in four groups (niche players, challengers, visionaries, and leaders) along two dimensions: completeness of vision (x-axis) and ability to execute (y-axis). As the quadrant clearly shows, most of the well-known BI/BA (business analytics) providers are positioned in the "leaders" category while many of the less known, relatively new, emerging providers are positioned in the "niche players" category.

The BI and analytics platform market's multiyear shift from IT-led enterprise reporting to business-led self-service analytics seems to have passed the tipping point. Most new buying is of modern, business-user-centric visual analytics platforms forcing a new market perspective, significantly reordering the vendor landscape. Most of the activity in the BI and analytics platform market is from organizations that are trying to mature their visualization capabilities and to move from descriptive to predictive and prescriptive analytics echelons. The vendors in the market have overwhelmingly concentrated on meeting this user demand. If there were a single market theme in 2015, it would be that data discovery/visualization became a mainstream architecture. While data discovery/visualization vendors such as Tableau, Qlik, and Microsoft are solidifying their position in the *Leaders* quadrant, others (both emerging and large, well-established tool/solution providers) are trying to move out of *Visionaries* into the *Leaders* quadrant.

This emphasis on data discovery/visualization from most of the leaders and visionaries in the market—which are now promoting tools with business-user-friendly data integration coupled with embedded storage and computing layers and unfettered drilling—continues to accelerate the trend toward decentralization and user empowerment of BI and analytics and greatly enables organizations' ability to perform diagnostic analytics.

Source: Gartner Magic Quadrant, released on February 4, 2016, **gartner.com** (accessed August 2016). Used with permission from Gartner Inc.

minimum, these warehouses involve a range of dimensions (e.g., product, location, organizational structure, time), a range of measures, and millions of cells of data. In an effort to address these challenges, a number of researchers have developed a variety of new visualization techniques.

Visual Analytics

Visual analytics is a recently coined term that is often used loosely to mean nothing more than information visualization. What is meant by **visual analytics** is the combination of visualization and predictive analytics. Whereas information visualization is aimed at answering "What happened?" and "What is happening?" and is closely associated with BI (routine reports, scorecards, and dashboards), visual analytics is aimed at answering "Why is it happening?" and "What is more likely to happen?" and is usually associated with business analytics (forecasting, segmentation, correlation analysis). Many of the information visualization vendors are adding the capabilities to call themselves visual analytics solution providers. One of the top, long-time analytics solution providers, SAS Institute, is approaching it from another direction. It is embedding its analytics capabilities into a high-performance data visualization environment that it calls *visual analytics*.

Visual or not visual, automated or manual, online or paper based, business reporting is not much different than telling a story. Technology Insights 3.3 provides a different, unorthodox viewpoint on better business reporting.

TECHNOLOGY INSIGHTS 3.3 Telling Great Stories with Data and Visualization

Everyone who has data to analyze has stories to tell, whether it's diagnosing the reasons for manufacturing defects, selling a new idea in a way that captures the imagination of your target audience, or informing colleagues about a particular customer service improvement program. And when it's telling the story behind a big strategic choice so that you and your senior management team can make a solid decision, providing a fact-based story can be especially challenging. In all cases, it's a big job. You want to be interesting and memorable; you know you need to keep it simple for your busy executives and colleagues. Yet you also know you have to be factual, detail oriented, and data driven, especially in today's metric-centric world.

It's tempting to present just the data and facts, but when colleagues and senior management are overwhelmed by data and facts without context, you lose. We have all experienced presentations with large slide decks only to find that the audience is so overwhelmed with data that they don't know what to think, or they are so completely tuned out that they take away only a fraction of the key points.

Start engaging your executive team and explaining your strategies and results more powerfully by approaching your assignment as a story. You will need the "what" of your story (the facts and data), but you also need the "Who?" "How?" "Why?" and the often-missed "So what?" It's these story elements that will make your data relevant and tangible for your audience. Creating a good story can aid you and senior management in focusing on what is important.

Why Story?

Stories bring life to data and facts. They can help you make sense and order out of a disparate collection of facts. They make it easier to remember key points and can paint a vivid picture of what the future can look like. Stories also create interactivity—people put themselves into stories and can relate to the situation.

Cultures have long used **storytelling** to pass on knowledge and content. In some cultures, storytelling is critical to their identity. For example, in New Zealand, some of the Maori people tattoo their faces with *mokus*. A *moku* is a facial tattoo containing a story about ancestors—the family tribe. A man may have a tattoo design on his face that shows features of a hammerhead to highlight unique qualities about his lineage. The design he chooses signifies what is part of his "true self" and his ancestral home.

Likewise, when we are trying to understand a story, the storyteller navigates to finding the "true north." If senior management is looking to discuss how they will respond to a competitive change, a good story can make sense and order out of a lot of noise. For example, you may have facts and data from two studies, one including results from an advertising study and one from a product satisfaction study. Developing a story for what you measured across both studies can help people see the whole where there were disparate parts. For rallying your distributors around a new product, you can employ a story to give vision to what the future can look like. Most important, storytelling is interactive—typically, the presenter uses words and pictures that audience members can put themselves into. As a result, they become more engaged and better understand the information.

So What Is a Good Story?

Most people can easily rattle off their favorite film or book. Or they remember a funny story that a colleague recently shared. Why do people remember these stories? Because they contain certain characteristics. First, a good story has great characters. In some cases, the reader or viewer has a vicarious experience where they become involved with the character. The character then has to be faced with a challenge that is difficult but believable. There must be hurdles that the character overcomes. And finally, the outcome or prognosis is clear by the end of the story. The situation may not be resolved—but the story has a clear endpoint.

Think of Your Analysis as a Story—Use a Story Structure

When crafting a data-rich story, the first objective is to find the story. Who are the characters? What is the drama or challenge? What hurdles have to be overcome? And at the end of your story, what do you want your audience to do as a result?

Once you know the core story, craft your other story elements: define your characters, understand the challenge, identify the hurdles, and crystallize the outcome or decision question. Make sure you are clear with what you want people to do as a result. This will shape how your audience will recall your story. With the story elements in place, write out the storyboard, which represents the structure and form of your story. Although it's tempting to skip this step, it is better first to understand the story you are telling and then to focus on the presentation structure and form. Once the storyboard is in place, the other elements will fall into place. The storyboard will help you think about the best analogies or metaphors, clearly set up challenge or opportunity, and finally see the flow and transitions needed. The storyboard also helps you focus on key visuals (graphs, charts, and graphics) that you need your executives to recall. Figure 3.24 shows a storyline for the impact of small loans in a worldwide view within the Tableau visual analytics environment.

In summary, do not be afraid to use data to tell great stories. Being factual, detail oriented, and data driven is critical in today's metric-centric world, but it does not have to mean being boring and lengthy. In fact, by finding the real stories in your data and following the best practices, you can get people to focus on your message—and thus on what's important. Here are those best practices:

1. Think of your analysis as a story—use a story structure.
2. Be authentic—your story will flow.
3. Be visual—think of yourself as a film editor.
4. Make it easy for your audience and you.
5. Invite and direct discussion.

Source: Fink, E., & Moore, S. J. (2012). "Five Best Practices for Telling Great Stories with Data." White paper by Tableau Software, Inc., **www.tableau.com/whitepapers/telling-data-stories** (accessed May 2016).

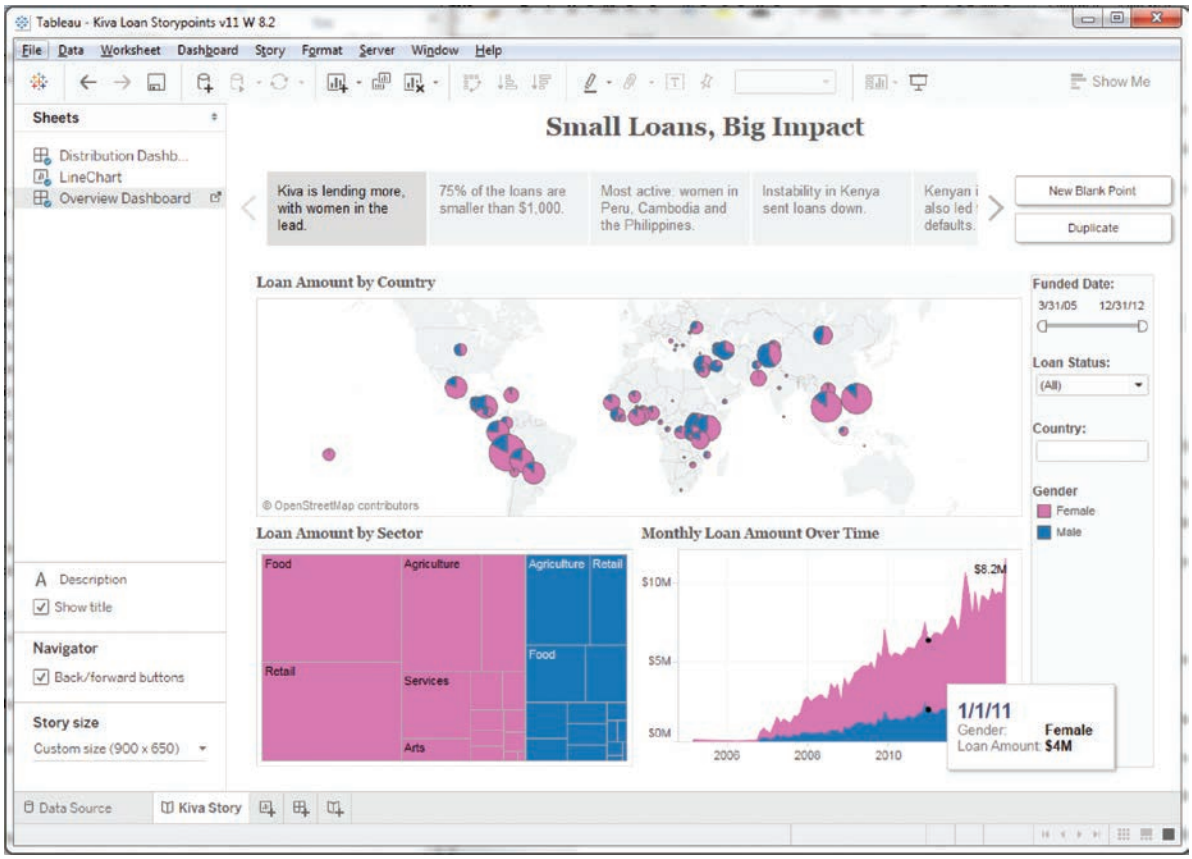


FIGURE 3.24 **A Storyline Visualization in Tableau Software.** *Source:* Used with permission from Tableau Software, Inc.

High-Powered Visual Analytics Environments

Due to the increasing demand for visual analytics coupled with fast-growing data volumes, there is an exponential movement toward investing in highly efficient visualization systems. With its latest move into visual analytics, the statistical software giant SAS Institute is now among those who are leading this wave. Its new product, SAS Visual Analytics, is a very **high-performance computing**, in-memory solution for exploring massive amounts of data in a very short time (almost instantaneously). It empowers users to spot patterns, identify opportunities for further analysis, and convey visual results via Web reports or mobile platforms such as tablets and smartphones. Figure 3.24 shows the high-level architecture of the SAS Visual Analytics platform. On one end of the architecture, there are universal data builder and administrator capabilities, leading into explorer, report designer, and mobile BI modules, collectively providing an end-to-end visual analytics solution.

Some of the key benefits proposed by the SAS analytics platform (see Figure 3.25) are the following:

- Empowers all users with data exploration techniques and approachable analytics to drive improved decision making. SAS Visual Analytics enables different types of users to conduct fast, thorough explorations on all available data. Sampling to reduce the data is not required and not preferred.
- Has easy-to-use, interactive Web interfaces that broaden the audience for analytics, enabling everyone to glean new insights. Users can look at additional options, make more precise decisions, and drive success even faster than before.

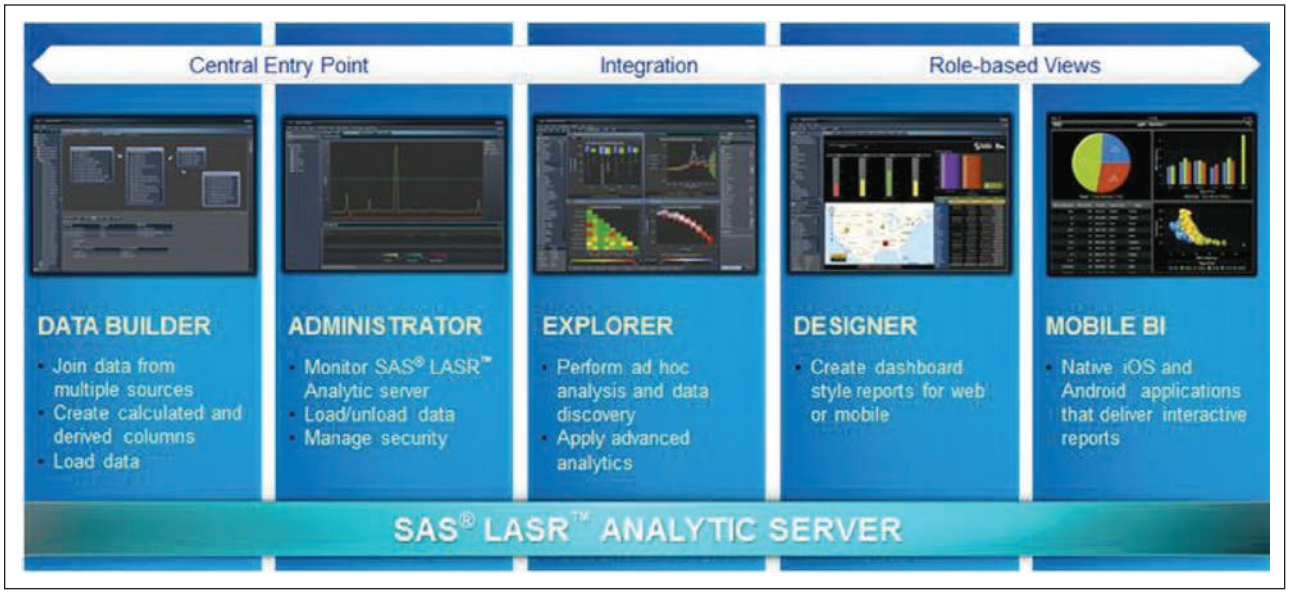


FIGURE 3.25 An Overview of SAS Visual Analytics Architecture. *Source:* Copyright © SAS Institute, Inc. Used with permission.

- Answers complex questions faster, enhancing the contributions from your analytic talent. SAS Visual Analytics augments the data discovery and exploration process by providing extremely fast results to enable better, more focused analysis. Analytically savvy users can identify areas of opportunity or concern from vast amounts of data so further investigation can take place quickly.
- Improves information sharing and collaboration. Large numbers of users, including those with limited analytical skills, can quickly view and interact with reports and charts via the Web, Adobe PDF files, and iPad mobile devices while IT maintains control of the underlying data and security. SAS Visual Analytics provides the right information to the right person at the right time to improve productivity and organizational knowledge.
- Liberates IT by giving users a new way to access the information they need. Frees IT from the constant barrage of demands from users who need access to different amounts of data, different data views, ad hoc reports, and one-off requests for information. SAS Visual Analytics enables IT to easily load and prepare data for multiple users. Once data are loaded and available, users can dynamically explore data, create reports, and share information on their own.
- Provides room to grow at a self-determined pace. SAS Visual Analytics provides the option of using commodity hardware or database appliances from EMC Greenplum and Teradata. It is designed from the ground up for performance optimization and scalability to meet the needs of any size organization.

Figure 3.26 shows a screenshot of a SAS Analytics platform where time-series forecasting and confidence interval around the forecast are depicted.

SECTION 3.10 REVIEW QUESTIONS

1. What are the main reasons for the recent emergence of visual analytics?
2. Look at Gartner's Magic Quadrant for Business Intelligence and Analytics Platforms. What do you see? Discuss and justify your observations.
3. What is the difference between information visualization and visual analytics?

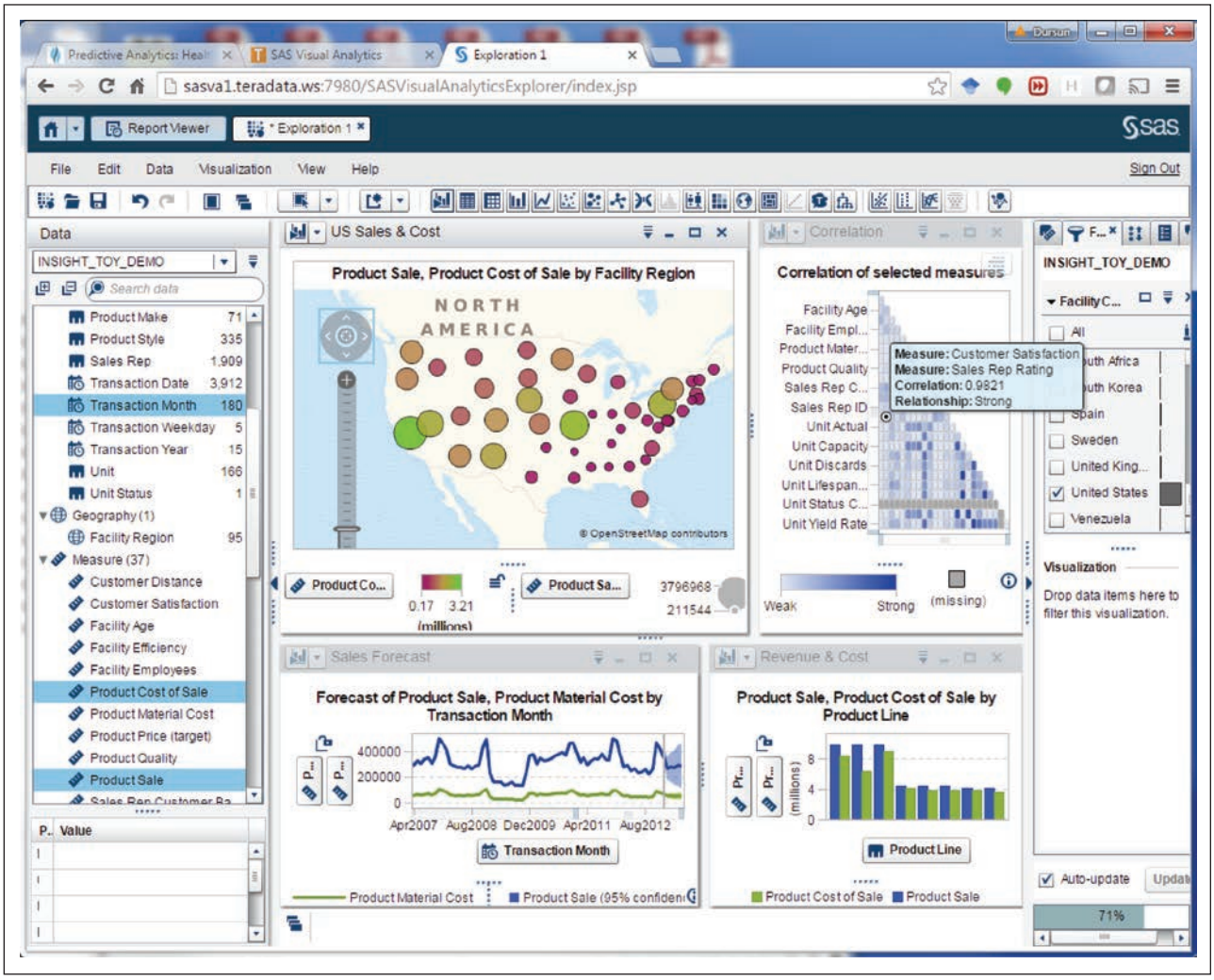


FIGURE 3.26 **A Screenshot from SAS Visual Analytics.** *Source:* Copyright © SAS Institute, Inc. Used with permission.

4. Why should storytelling be a part of your reporting and data visualization?
5. What is a high-powered visual analytics environment? Why do we need it?

3.11 INFORMATION DASHBOARDS

Information dashboards are common components of most, if not all, BI or business analytics platforms, business performance management systems, and performance measurement software suites. **Dashboards** provide visual displays of important information that is consolidated and arranged on a single screen so that the information can be digested at a single glance and easily drilled in and further explored. A typical dashboard is shown in Figure 3.27. This particular executive dashboard displays a variety of key performance indicators (KPIs) for a hypothetical software company called Sonatica (selling audio tools). This executive dashboard shows a high-level view of the different functional groups surrounding the products, starting from a general overview to the marketing efforts, sales, finance, and support departments. All of this is intended to give executive

FIGURE 3.27 **A Sample Executive Dashboard.** *Source:* A Sample Executive Dashboard from Dundas Data Visualization, Inc., **www.dundas.com,** reprinted with permission.

decision makers a quick and accurate idea of what is going on within the organization. On the left side of the dashboard, we can see (in a time-series fashion) the quarterly changes in revenues, expenses, and margins as well as the comparison of those figures to previous years' monthly numbers. On the upper-right side are two dials with color-coded regions showing the amount of monthly expenses for support services (dial on the left) and the amount of other expenses (dial on the right). As the color coding indicates, although the monthly support expenses are well within the normal ranges, the other expenses are in the red region, indicating excessive values. The geographic map on the bottom right shows the distribution of sales at the country level throughout the world. Behind these graphical icons there are various mathematical functions aggregating numerous data points to their highest level of meaningful figures. By clicking on these graphical icons, the consumer of this information can drill down to more granular levels of information and data.

Dashboards are used in a wide variety of businesses for a wide variety of reasons. For instance, in Application Case 3.7, you will find the summary of a successful implementation of information dashboards by the Dallas Cowboys football team.

Application Case 3.7 Dallas Cowboys Score Big with Tableau and Teknion

Founded in 1960, the Dallas Cowboys are a professional American football team headquartered in Irving, Texas. The team has a large national following, which is perhaps best represented by their NFL record for number of consecutive games at sold-out stadiums.

The Challenge

Bill Priakos, chief operating officer (COO) of the Dallas Cowboys Merchandising Division, and his team needed more visibility into their data so they could run it more profitably. Microsoft was selected as the baseline platform for this upgrade as well as a number of other sales, logistics, and e-commerce (per MW) applications. The Cowboys expected that this new information architecture would provide the needed analytics and reporting. Unfortunately, this was not the case, and the search began for a robust dashboarding, analytics, and reporting tool to fill this gap.

The Solution and Results

Tableau and Teknion together provided real-time reporting and dashboard capabilities that exceeded the Cowboys' requirements. Systematically and methodically, the Teknion team worked side by side with data owners and data users within the Dallas Cowboys to deliver all required functionality on time and under budget. "Early in the process, we were able to get a clear understanding of what it would take to run a more profitable operation for the Cowboys," said Teknion Vice President Bill Luisi. "This process step is a key step in Teknion's approach with any client, and it always pays huge dividends as the implementation plan progresses." Added Luisi, "Of course, Tableau worked very closely with us and the Cowboys during the entire project. Together, we made sure that the Cowboys could achieve their reporting and analytical goals in record time."

Now, for the first time, the Dallas Cowboys are able to monitor their complete merchandising activities from manufacture to end customer and not only see what is happening across the life cycle but also drill down even further into why it is happening.

Today, this BI solution is used to report and analyze the business activities of the Merchandising Division, which is responsible for all of the Dallas Cowboys' brand sales. Industry estimates say that the Cowboys generate 20 percent of all NFL merchandise sales, which reflects the fact that they are the most recognized sports franchise in the world.

According to Eric Lai, a *ComputerWorld* reporter, Tony Romo and the rest of the Dallas Cowboys may have been only average on the football field in the last few years, but off the field, especially in the merchandising arena, they remain America's team.

Questions for Case 3.7

1. How did the Dallas Cowboys use information visualization?
2. What were the challenge, the proposed solution, and the obtained results?

Sources: Lai, E. (2009, October 8). "BI Visualization Tool Helps Dallas Cowboys Sell More Tony Romo Jerseys," *ComputerWorld*. Tableau case study. **tableau.com/blog/computerworld-dallas-cowboys-business-intelligence** (accessed July 2018).

Dashboard Design

Dashboards are not a new concept. Their roots can be traced at least to the executive information system of the 1980s. Today, dashboards are ubiquitous. For example, a few years back, Forrester Research estimated that over 40 percent of the largest 2,000 companies in the world used the technology (Ante & McGregor, 2006). Since then, one can safely assume that this number has gone up quite significantly. In fact, today it would be rather unusual to see a large company using a BI system that does not employ some sort of performance dashboards. The Dashboard Spy Web site (**dashboardspy.com/about**) provides further evidence of their ubiquity. The site contains descriptions and screenshots

of thousands of BI dashboards, scorecards, and BI interfaces used by businesses of all sizes and industries, nonprofits, and government agencies.

According to Eckerson (2006), a well-known expert on BI in general and dashboards in particular, the most distinctive feature of a dashboard is its three layers of information:

1. ***Monitoring:*** Graphical, abstracted data to monitor key performance metrics.
2. ***Analysis:*** Summarized dimensional data to analyze the root cause of problems.
3. ***Management:*** Detailed operational data that identify what actions to take to resolve a problem.

Because of these layers, dashboards pack a large amount of information into a single screen. According to Few (2005), "The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly." To speed assimilation of the numbers, they need to be placed in context. This can be done by comparing the numbers of interest to other baseline or target numbers, by indicating whether the numbers are good or bad, by denoting whether a trend is better or worse, and by using specialized display widgets or components to set the comparative and evaluative context. Some of the common comparisons that are typically made in BI systems include comparisons against past values, forecasted values, targeted values, benchmark or average values, multiple instances of the same measure, and the values of other measures (e.g., revenues versus costs).

Even with comparative measures, it is important to specifically point out whether a particular number is good or bad and whether it is trending in the right direction. Without these types of evaluative designations, it can be time consuming to determine the status of a particular number or result. Typically, either specialized visual objects (e.g., traffic lights, dials, and gauges) or visual attributes (e.g., color coding) are used to set the evaluative context. An interactive dashboard-driven reporting data exploration solution built by an energy company is featured in Application Case 3.8.

Application Case 3.8 Visual Analytics Helps Energy Supplier Make Better Connections

Energy markets all around the world are going through a significant change and transformation, creating ample opportunities along with significant challenges. As is the case in any industry, opportunities are attracting more players in the marketplace, increasing the competition, and reducing the tolerances for less-than-optimal business decision making. Success requires creating and disseminating accurate and timely information to whomever whenever it is needed. For instance, if you need to easily track marketing budgets, balance employee workloads, and target customers with tailored marketing messages, you would need three different reporting solutions. Electrabel GDF SUEZ is doing all of that for its marketing and sales business unit with SAS®Analytics Visual Analytics platform.

The one-solution approach is a great time-saver for marketing professionals in an industry that is undergoing tremendous change. "It is a huge challenge to stabilize our market position in the energy market. That includes volume, prices, and margins for both retail and business customers," notes Danny Noppe, manager of Reporting Architecture and Development in the Electrabel Marketing and Sales business unit. The company is the largest supplier of electricity in Belgium and the largest producer of electricity for Belgium and the Netherlands. Noppe says it is critical that Electrabel increase the efficiency of its customer communications as it explores new digital channels and develops new energy-related services.

"The better we know the customer, the better our likelihood of success," he says. "That is why

(*Continued*)

Application Case 3.8 (Continued)

we combine information from various sources—phone traffic with the customer, online questions, text messages, and mail campaigns. This enhanced knowledge of our customer and prospect base will be an additional advantage within our competitive market."

One Version of the Truth

Electrabel was using various platforms and tools for reporting purposes. This sometimes led to ambiguity in the reported figures. The utility also had performance issues in processing large data volumes. SAS Visual Analytics with in-memory technology removes the ambiguity and the performance issues. "We have the autonomy and flexibility to respond to the need for customer insight and data visualization internally," Noppe says. "After all, fast reporting is an essential requirement for action-oriented departments such as sales and marketing."

Working More Efficiently at a Lower Cost

SAS Visual Analytics automates the process of updating information in reports. Instead of building a report that is out of date by the time it is completed, the data are refreshed for all the reports once a week and is available on dashboards. In deploying the solution, Electrabel chose a phased approach, starting with simple reports and moving on to more complex ones. The first report took a few weeks to build, and the rest came quickly. The successes include the following:

- Reduction of data preparation from two days to only two hours.
- Clear graphic insight into the invoicing and composition of invoices for business-to-business (B2B) customers.
- A workload management report by the operational teams. Managers can evaluate team workloads on a weekly or long-term basis and can make adjustments accordingly.

"We have significantly improved our efficiency and can deliver quality data and reports more frequently, and at a significantly lower cost," says Noppe. And if the company needs to combine data from multiple sources, the process is equally easy. "Building visual reports, based on these data marts, can be achieved in a few days, or even a few hours."

Noppe says the company plans to continue broadening its insight into the digital behavior of its customers, combining data from Web analytics, e-mail, and social media with data from back-end systems. "Eventually, we want to replace all labor-intensive reporting with SAS Visual Analytics," he says, adding that the flexibility of SAS Visual Analytics is critical for his department. "This will give us more time to tackle other challenges. We also want to make this tool available on our mobile devices. This will allow our account managers to use up-to-date, insightful, and adaptable reports when visiting customers. We've got a future-oriented reporting platform to do all we need."

Questions for Case 3.8

1. Why do you think energy supply companies are among the prime users of information visualization tools?
2. How did Electrabel use information visualization for the single version of the truth?
3. What were their challenges, the proposed solution, and the obtained results?

Source: SAS Customer Story, "Visual Analytics Helps Energy Supplier Make Better Connections." **http://www.sas.com/en_us/customers/electrabel-be.html** (accessed July 2018).

What to Look for in a Dashboard

Although performance dashboards and other information visualization frameworks differ, they all share some common design characteristics. First, they all fit within the larger BI and/or performance measurement system. This means that their underlying architecture is the BI or performance management architecture of the larger system. Second, all well-designed dashboards and other information visualizations possess the following characteristics (Novell, 2009):

- They use visual components (e.g., charts, performance bars, sparklines, gauges, meters, stoplights) to highlight, at a glance, the data and exceptions that require action.
- They are transparent to the user, meaning that they require minimal training and are extremely easy to use.
- They combine data from a variety of systems into a single, summarized, unified view of the business.
- They enable drill-down or drill-through to underlying data sources or reports, providing more detail about the underlying comparative and evaluative context.
- They present a dynamic, real-world view with timely data refreshes, enabling the end user to stay up-to-date with any recent changes in the business.
- They require little, if any, customized coding to implement, deploy, and maintain.

Best Practices in Dashboard Design

The real estate saying "location, location, location" makes it obvious that the most important attribute for a piece of real estate property is where it is located. For dashboards, it is "data, data, data." Often overlooked, data are considered one of the most important things to focus on in designing dashboards (Carotenuto, 2007). Even if a dashboard's appearance looks professional, is aesthetically pleasing, and includes graphs and tables created according to accepted visual design standards, it is also important to ask about the data: Are they reliable? Are they timely? Are any data missing? Are they consistent across all dashboards? Here are some of the experience-driven best practices in dashboard design (Radha, 2008).

Benchmark Key Performance Indicators with Industry Standards

Many customers, at some point in time, want to know if the metrics they are measuring are the right metrics to monitor. Sometimes customers have found that the metrics they are tracking are not the right ones to track. Doing a gap assessment with industry benchmarks aligns you with industry best practices.

Wrap the Dashboard Metrics with Contextual Metadata

Often when a report or a visual dashboard/scorecard is presented to business users, questions remain unanswered. The following are some examples:

- Where did you source these data?
- While loading the data warehouse, what percentage of the data was rejected/encountered data quality problems?
- Is the dashboard presenting "fresh" information or "stale" information?
- When was the data warehouse last refreshed?
- When is it going to be refreshed next?
- Were any high-value transactions that would skew the overall trends rejected as a part of the loading process?

Validate the Dashboard Design by a Usability Specialist

In most dashboard environments, the dashboard is designed by a tool specialist without giving consideration to usability principles. Even though it is a well-engineered data warehouse that can perform well, many business users do not use the dashboard because it is perceived as not being user friendly, leading to poor adoption of the infrastructure and change management issues. Up-front validation of the dashboard design by a usability specialist can mitigate this risk.

Prioritize and Rank Alerts/Exceptions Streamed to the Dashboard

Because there are tons of raw data, having a mechanism by which important exceptions/behaviors are proactively pushed to the information consumers is important. A business rule can be codified, which detects the alert pattern of interest. It can be coded into a program, using database-stored procedures, which can crawl through the fact tables and detect patterns that need immediate attention. This way, information finds the business user as opposed to the business user polling the fact tables for the occurrence of critical patterns.

Enrich the Dashboard with Business-User Comments

When the same dashboard information is presented to multiple business users, a small text box can be provided that can capture the comments from an end user's perspective. This can often be tagged to the dashboard to put the information in context, adding perspective to the structured KPIs being rendered.

Present Information in Three Different Levels

Information can be presented in three layers depending on the granularity of the information: the visual dashboard level, the static report level, and the self-service cube level. When a user navigates the dashboard, a simple set of 8 to 12 KPIs can be presented, which would give a sense of what is going well and what is not.

Pick the Right Visual Construct Using Dashboard Design Principles

In presenting information in a dashboard, some information is presented best with bar charts and some with time-series line graphs, and when presenting correlations, a scatter plot is useful. Sometimes merely rendering it as simple tables is effective. Once the dashboard design principles are explicitly documented, all the developers working on the front end can adhere to the same principles while rendering the reports and dashboard.

Provide for Guided Analytics

In a typical organization, business users can be at various levels of analytical maturity. The capability of the dashboard can be used to guide the "average" business user to access the same navigational path as that of an analytically savvy business user.

SECTION 3.11 REVIEW QUESTIONS

1. What is an information dashboard? Why is it so popular?
2. What are the graphical widgets commonly used in dashboards? Why?
3. List and describe the three layers of information portrayed on dashboards.
4. What are the common characteristics of dashboards and other information visuals?
5. What are the best practices in dashboard design?

Chapter Highlights

- Data have become one of the most valuable assets of today's organizations.
- Data are the main ingredient for any BI, data science, and business analytics initiative.
- Although its value proposition is undeniable, to live up its promise, the data must comply with some basic usability and quality metrics.

- The term *data* (*datum* in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences.
- At the highest level of abstraction, data can be classified as structured and unstructured.
- Data in original/raw form are not usually ready to be useful in analytics tasks.
- Data preprocessing is a tedious, time-demanding, yet crucial task in business analytics.
- Statistics is a collection of mathematical techniques to characterize and interpret data.
- Statistical methods can be classified as either descriptive or inferential.
- Statistics in general, as well as descriptive statistics in particular, is a critical part of BI and business analytics.
- Descriptive statistics methods can be used to measure central tendency, dispersion, or the shape of a given data set.
- Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics.
- Linear regression and logistic regression are the two major regression types in statistics.
- Logistics regression is a probability-based classification algorithm.
- Time series is a sequence of data points of a variable, measured and recorded at successive points in time spaced at uniform time intervals.
- A report is any communication artifact prepared with the specific intention of conveying information in a presentable form.
- A business report is a written document that contains information regarding business matters.
- The key to any successful business report is clarity, brevity, completeness, and correctness.
- Data visualization is the use of visual representations to explore, make sense of, and communicate data.
- Perhaps the most notable information graphic of the past was developed by Charles J. Minard, who graphically portrayed the losses suffered by Napoleon's army in the Russian campaign of 1812.
- Basic chart types include line, bar, and pie chart.
- Specialized charts are often derived from the basic charts as exceptional cases.
- Data visualization techniques and tools make the users of business analytics and BI systems better information consumers.
- Visual analytics is the combination of visualization and predictive analytics.
- Increasing demand for visual analytics coupled with fast-growing data volumes led to exponential growth in highly efficient visualization systems investment.
- Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored.

Key terms

analytics ready
arithmetic mean
box-and-whiskers plot
box plot
bubble chart
business report
categorical data
centrality
correlation
dashboards
data preprocessing
data quality
data security
data taxonomy
data visualization
datum
descriptive statistics
dimensional reduction
dispersion
high-performance computing
histogram
inferential statistics
key performance indicator (KPI)
knowledge
kurtosis
learning
linear regression
logistic regression
mean absolute deviation
median
mode
nominal data
online analytics processing (OLAP)
ordinal data
ordinary least squares (OLS)
pie chart
quartile
range
ratio data
regression
report
scatter plot
skewness
standard deviation
statistics
storytelling
structured data
time-series forecasting
unstructured data
variable selection
variance
visual analytics

Questions for Discussion

1. How do you describe the importance of data in analytics? Can we think of analytics without data? Explain.
2. Considering the new and broad definition of business analytics, what are the main inputs and outputs to the analytics continuum?
3. Where do the data for business analytics come from? What are the sources and the nature of those incoming data?
4. What are the most common metrics that make for analytics-ready data?
5. What are the main categories of data? What types of data can we use for BI and analytics?
6. Can we use the same data representation for all analytics models (i.e., do different analytics models require different data representation schema)? Why, or why not?
7. Why are the original/raw data not readily usable by analytics tasks?
8. What are the main data preprocessing steps? List and explain their importance in analytics.
9. What does it mean to clean/scrub the data? What activities are performed in this phase?
10. Data reduction can be applied to rows (sampling) and/or columns (variable selection). Which is more challenging? Explain.
11. What is the relationship between statistics and business analytics? (Consider the placement of statistics in a business analytics taxonomy.)
12. What are the main differences between descriptive and inferential statistics?
13. What is a box-and-whiskers plot? What types of statistical information does it represent?
14. What are the two most commonly used shape characteristics to describe a data distribution?
15. List and briefly define the central tendency measures of descriptive statistics.
16. What are the commonalities and differences between regression and correlation?
17. List and describe the main steps to follow in developing a linear regression model.
18. What are the most commonly pronounced assumptions for linear regression? What is crucial to the regression models against these assumptions?
19. What are the commonalities and differences between linear regression and logistic regression?
20. What is time series? What are the main forecasting techniques for time-series data?
21. What is a business report? Why is it needed?
22. What are the best practices in business reporting? How can we make our reports stand out?
23. Describe the cyclic process of management, and comment on the role of business reports.
24. List and describe the three major categories of business reports.
25. Why has information visualization become a centerpiece in BI and business analytics? Is there a difference between information visualization and visual analytics?
26. What are the main types of charts/graphs? Why are there so many of them?
27. How do you determine the right chart for a job? Explain and defend your reasoning.
28. What is the difference between information visualization and visual analytics?
29. Why should storytelling be a part of your reporting and data visualization?
30. What is an information dashboard? What does it present?
31. What are the best practices in designing highly informative dashboards?
32. Do you think information/performance dashboards are here to stay? Or are they about to be outdated? What do you think will be the next big wave in BI and business analytics in terms of data/information visualization?

Exercises

Teradata University and Other Hands-on Exercises

1. Download the "Voting Behavior" data and the brief data description from the book's Web site. This is a data set manually compiled from counties all around the United States. The data are partially processed, that is, some derived variables have been created. Your task is to thoroughly preprocess the data by identifying the error and anomalies and proposing remedies and solutions. At the end, you should have an analytics-ready version of these data. Once the preprocessing is completed, pull these data into Tableau (or into some other data visualization software tool) to extract useful visual information from it. To do so, conceptualize relevant questions and hypotheses (come up with at least three of them) and create proper visualizations that address those questions of "tests" of those hypotheses.
2. Download Tableau (at **tableau.com**, following academic free software download instructions on the site). Using the Visualization_MFG_Sample data set (available as an Excel file on this book's Web site), answer the following questions:
   a. What is the relationship between gross box office revenue and other movie-related parameters given in the data set?

   b. How does this relationship vary across different years? Prepare a professional-looking written report that is enhanced with screenshots of your graphic findings.
3. Go to **teradatauniversitynetwork.com**. Look for an article that deals with the nature of data, management of data, and/or governance of data as it relates to BI and analytics, and critically analyze the content of the article.
4. Go to UCI data repository (**archive.ics.uci.edu/ml/datasets.html**) and identify a large data set that contains both numeric and nominal values. Using Microsoft Excel or any other statistical software:
   a. Calculate and interpret central tendency measures for each and every variable.
   b. Calculate and interpret the dispersion/spread measures for each and every variable.
5. Go to UCI data repository (**archive.ics.uci.edu/ml/datasets.html**) and identify two data sets, one for estimation/regression and one for classification. Using Microsoft Excel or any other statistical software:
   a. Develop and interpret a linear regression model.
   b. Develop and interpret a logistic regression model.
6. Go to **KDnuggest.com** and become familiar with the range of analytics resources available on this portal. Then identify an article, a white paper, or an interview script that deals with the nature of data, management of data, and/or governance of data as they relate to BI and business analytics, and critically analyze the content of the article.
7. Go to Stephen Few's blog, "The Perceptual Edge" (**perceptualedge.com**). Go to the section of "Examples." In this section, he provides critiques of various dashboard examples. Read a handful of these examples. Now go to **dundas.com**. Select the "Gallery" section of the site. Once there, click the "Digital Dashboard" selection. You will be shown a variety of different dashboard demos. Run a couple of them.
   a. What types of information and metrics are shown on the demos? What types of actions can you take?
   b. Using some of the basic concepts from Few's critiques, describe some of the good design points and bad design points of the demos.
8. Download an information visualization tool, such as Tableau, QlikView, or Spotfire. If your school does not have an educational agreement with these companies, a trial version would be sufficient for this exercise. Use your own data (if you have any) or use one of the data sets that comes with the tool (such tools usually have one or more data sets for demonstration purposes). Study the data, come up with several business problems, and use data visualization to analyze, visualize, and potentially solve those problems.
9. Go to **teradatauniversitynetwork.com**. Find the "Tableau Software Project." Read the description, execute the tasks, and answer the questions.
10. Go to **teradatauniversitynetwork.com**. Find the assignments for SAS Visual Analytics. Using the information and step-by-step instructions provided in the assignment, execute the analysis on the SAS Visual Analytics tool (which is a Web-enabled system that does not require any local installation). Answer the questions posed in the assignment.
11. Find at least two articles (one journal article and one white paper) that talk about storytelling, especially within the context of analytics (i.e., data-driven storytelling). Read and critically analyze the article and paper, and write a report to reflect your understanding and opinions about the importance of storytelling in BI and business analytics.
12. Go to **data.gov**—a U.S. government–sponsored data portal that has a very large number of data sets on a wide variety of topics ranging from healthcare to education, climate to public safety. Pick a topic that you are most passionate about. Go through the topic-specific information and explanation provided on the site. Explore the possibilities of downloading the data, and use your favorite data visualization tool to create your own meaningful information and visualizations.

Team Assignments and Role-Playing Projects

1. Analytics starts with data. Identifying, accessing, obtaining, and processing of relevant data is the most essential task in any analytics study. As a team, you are tasked to find a large enough real-world data (either from your own organization, which is the most preferred, or from the Internet that can start with a simple search, or from the data links posted on **KDnuggets.com**), one that has tens of thousands of rows and more than 20 variables to go through, and document a thorough data preprocessing project. In your processing of the data, identify anomalies and discrepancies using descriptive statistics methods and measures, and make the data analytics ready. List and justify your preprocessing steps and decisions in a comprehensive report.
2. Go to a well-known information dashboard provider Web site (**dundas.com**, **idashboards.com**, **enterprisedashboard.com**). These sites provide a number of examples of executive dashboards. As a team, select a particular industry (e.g., healthcare, banking, airline). Locate a handful of example dashboards for that industry. Describe the types of metrics found on the dashboards. What types of displays are used to provide the information? Using what you know about dashboard design, provide a paper prototype of a dashboard for this information.
3. Go to **teradatauniversitynetwork.com**. From there, go to University of Arkansas data sources. Choose one of the large data sets, and download a large number of records (this could require you to write an SQL statement that creates the variables that you want to include in the data set). Come up with at least 10 questions that can be addressed with information visualization. Using your favorite data visualization tool (e.g., Tableau), analyze the data, and prepare a detailed report that includes screenshots and other visuals.

References

Abela, A. (2008). *Advanced Presentations by Design: Creating Communication That Drives Action*. New York, NY: Wiley.

Annas, G. (2003). "HIPAA Regulations—A New Era of Medical-Record Privacy?" *New England Journal of Medicine, 348*(15), 1486–1490.

Ante, S., & J. McGregor. (2006). "Giving the Boss the Big Picture: A Dashboard Pulls Up Everything the CEO Needs to Run the Show." *Business Week*, 43–51.

Carotenuto, D. (2007). "Business Intelligence Best Practices for Dashboard Design." WebFOCUS. **www.datawarehouse.inf.br/papers/information_builders_dashboard_best_practices.pdf** (accessed August 2016).

Dell Customer Case Study. "Medical Device Company Ensures Product Quality While Saving Hundreds of Thousands of Dollars." **https://software.dell.com/documents/instrumentation-laboratory-medical-device-companyensures-product-quality-while-saving-hundreds-ofthousands-of-dollars-case-study-80048.pdf** (accessed August 2016).

Delen, D. (2010). "A Comparative Analysis of Machine Learning Techniques for Student Retention Management." *Decision Support Systems, 49*(4), 498–506.

Delen, D. (2011). "Predicting Student Attrition with Data Mining Methods." *Journal of College Student Retention 13*(1), 17–35.

Delen, D. (2015). *Real-World Data Mining: Applied Business Analytics and Decision Making*. Upper Saddle River, NJ: Financial Times Press (A Pearson Company).

Delen, D., D. Cogdell, & N. Kasap. (2012). "A Comparative Analysis of Data Mining Methods in Predicting NCAA Bowl Outcomes." *International Journal of Forecasting, 28*, 543–552.

Eckerson, W. (2006). *Performance Dashboards*. New York: Wiley.

Few, S. (2005, Winter). "Dashboard Design: Beyond Meters, Gauges, and Traffic Lights." *Business Intelligence Journal, 10*(1).

Few, S. (2007). "Data Visualization: Past, Present and Future." **Perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf** (accessed July 2016).

Fink, E., & S. J. Moore. (2012). "Five Best Practices for Telling Great Stories with Data." Tableau Software, Inc. **www.tableau.com/whitepapers/telling-data-stories** (accessed May 2016).

Freeman, K., & R. M. Brewer. (2016). "The Politics of American College Football." *Journal of Applied Business and Economics, 18*(2), 97–101.

Gartner Magic Quadrant. (2016, February 4). **gartner.com** (accessed August 2016).

Grimes, S. (2009a, May 2). "Seeing Connections: Visualizations Makes Sense of Data. *Intelligent Enterprise*." **i.cmpnet.com/intelligententerprise/next-era-business-intelligence/Intelligent_Enterprise_Next_Era_BI_Visualization.pdf** (accessed January 2010).

Grimes, S. (2009b). Text "Analytics 2009: User Perspectives on Solutions and Providers." Alta Plana. **altaplana.com/TextAnalyticsPerspectives2009.pdf** (accessed July, 2016).

Hardin, M. Hom, R. Perez, & Williams L. (2012). "Which Chart or Graph Is Right for You?" Tableau Software. **http://www.tableau.com/sites/default/files/media/which_chart_v6_final_0.pdf** (accessed August 2016).

Hernández, M., & S. J. Stolfo. (1998, January). "Real-World Data Is Dirty: Data Cleansing and the Merge/Purge Problem." *Data Mining and Knowledge Discovery, 2*(1), 9–37.

Hill, G. (2016). "A Guide to Enterprise Reporting." **Ghill.customer.netspace.net.au/reporting/definition.html** (accessed July 2016).

Kim, W., B. J. Choi, E. K. Hong, S. K. Kim, & D. Lee. (2003). "A Taxonomy of Dirty Data." *Data Mining and Knowledge Discovery, 7*(1), 81–99.

Kock, N. F., R. J. McQueen, & J. L. Corner. (1997). "The Nature of Data, Information and Knowledge Exchanges in Business Processes: Implications for Process Improvement and Organizational Learning." *The Learning Organization, 4*(2), 70–80.

Kotsiantis, S., D. Kanellopoulos, & P. E. Pintelas. (2006). "Data Preprocessing for Supervised Leaning." *International Journal of Computer Science, 1*(2), 111–117.

Lai, E. (2009, October 8). "BI Visualization Tool Helps Dallas Cowboys Sell More Tony Romo Jerseys." *ComputerWorld*.

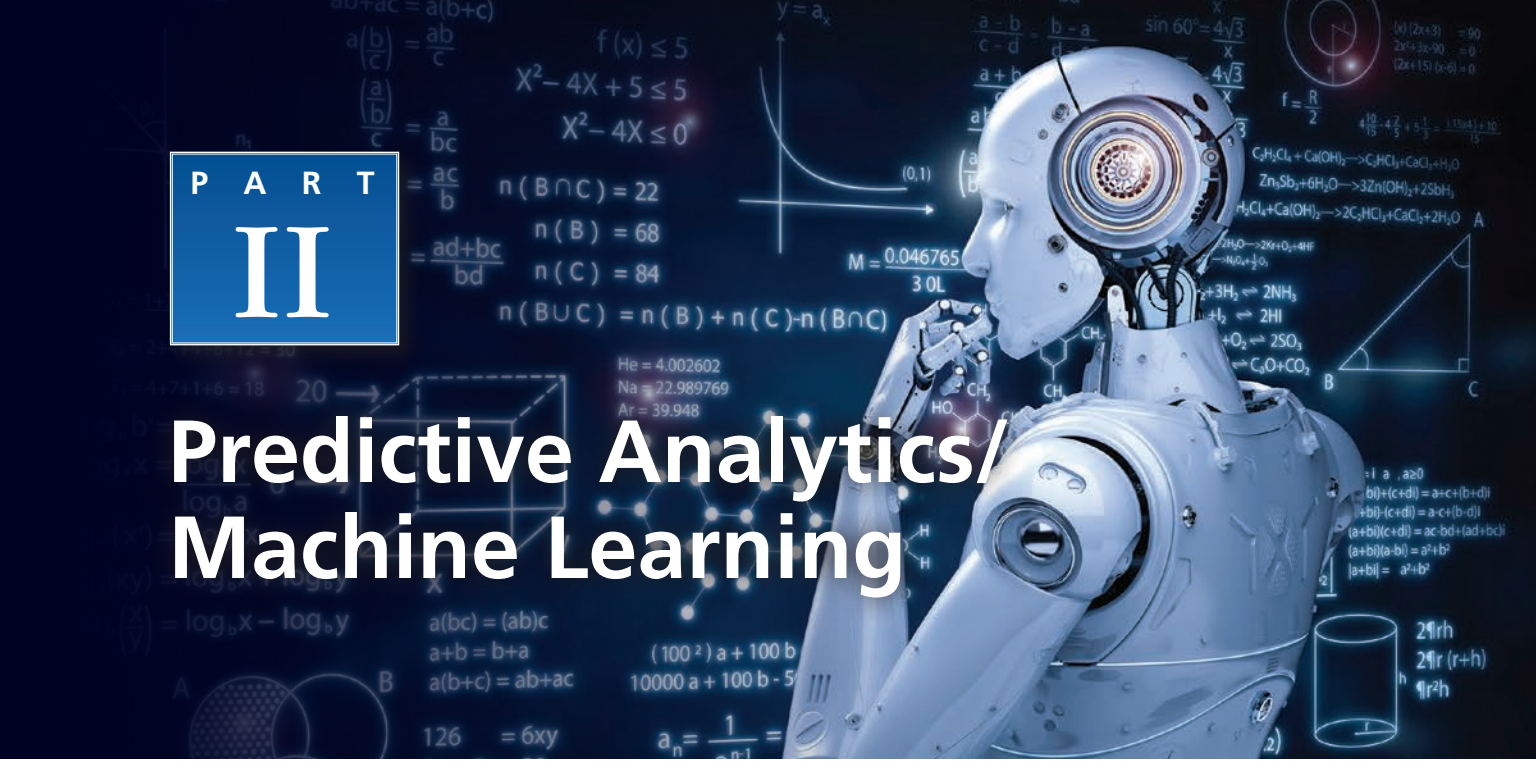
Quinn, C. (2016). "Data-Driven Marketing at SiriusXM," Teradata Articles & News. **http://bigdata.teradata.com/US/Articles-News/Data-Driven-Marketing-At-SiriusXM/** (accessed August 2016); "SiriusXM Attracts and Engages a New Generation of Radio Consumers." **http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB8597.pdf?processed=1** (accessed August 2018).

Novell. (2009, April). "Executive Dashboards Elements of Success." Novell white paper. **www.novell.com/docrep/documents/3rkw3etfc3/Executive%20Dashboards_Elements_of_Success_White_Paper_en.pdf** (accessed June 2016).

Radha, R. (2008). "Eight Best Practices in Dashboard Design." *Information Management*. **www.information-management.com/news/columns/-10001129-1.html** (accessed July 2016).

SAS. (2014). "Data Visualization Techniques: From Basics to Big Data." **http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-visualization-techniques-106006.pdf** (accessed July 2016).

Thammasiri, D., D. Delen, P. Meesad, & N. Kasap. (2014). "A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition." *Expert Systems with Applications, 41*(2), 321–330.



PART II

Predictive Analytics/ Machine Learning

CHAPTER

4

Data Mining Process, Methods, and Algorithms

LEARNING OBJECTIVES

- Define *data mining* as an enabling technology for business analytics
- Understand the objectives and benefits of data mining
- Become familiar with the wide range of applications of data mining
- Learn the standardized data mining processes
- Learn different methods and algorithms of data mining
- Build awareness of existing data mining software tools
- Understand the privacy issues, pitfalls, and myths of data mining

Generally speaking, data mining is a way to develop intelligence (i.e., actionable information or knowledge) from data that an organization collects, organizes, and stores. A wide range of data mining techniques is being used by organizations to gain a better understanding of their customers and their operations and to solve complex organizational problems. In this chapter, we study data mining as an enabling technology for business analytics and predictive analytics; learn about the standard processes of conducting data mining projects; understand and build expertise in the use of major data mining techniques; develop awareness of the existing software tools; and explore privacy issues, common myths, and pitfalls that are often associated with data mining.

4.1 OPENING VIGNETTE: Miami-Dade Police Department Is Using Predictive Analytics to Foresee and Fight Crime

Predictive analytics and data mining have become an integral part of many law enforcement agencies including the Miami-Dade Police Department whose mission is not only to protect the safety of Florida's largest county, with 2.5 million citizens (making it the seventh largest in the United States), but also to provide a safe and inviting climate for the millions of tourists who come from around the world to enjoy the county's natural beauty, warm climate, and stunning beaches. With tourists spending nearly US$20 billion every year and generating nearly one-third of Florida's sales taxes, it is hard to overstate the importance of tourism to the region's economy. So although few of the county's police officers would likely list economic development in their job description, nearly all grasp the vital link between safe streets and the region's tourist-driven prosperity.

That connection is paramount for Lieutenant Arnold Palmer, currently supervising the Robbery Investigations Section and a former supervisor of the department's Robbery Intervention Detail. This specialized team of detectives is focused on intensely policing the county's robbery hot spots and worst repeat offenders. He and the team occupy modest offices on the second floor of a modern-looking concrete building set back from a palm-lined street on the western edge of Miami. In his 10 years in the unit and 23 in total on the force, Palmer has seen many changes—not just in policing practices like the way his team used to mark street crime hot spots with colored pushpins on a map.

POLICING WITH LESS

Palmer and the team have also seen the impact of a growing population, shifting demographics, and a changing economy on the streets they patrol. Like any good police force officers, they have continually adapted their methods and practices to meet a policing challenge that has grown in scope and complexity. But like nearly all branches of the county's government, intensifying budget pressures have placed the department in a squeeze between rising demands and shrinking resources.

Palmer, who sees detectives as front-line fighters against a rising tide of street crime and the looming prospect of ever-tightening resources, put it this way: "Our basic challenge was how to cut street crime even as tighter resources have reduced the number of cops on the street." Over the years, the team has been open to trying new tools, the most notable of which is a program called "analysis-driven enforcement," which used crime history data as the basis for positioning teams of detectives. "We've evolved a lot since then in our ability to predict where robberies are likely to occur, both through the use of analysis and our own collective experience."

NEW THINKING ON COLD CASES

The more confounding challenge for Palmer and his team of investigators, one shared with the police of all major urban areas, is in closing the hardest cases whose leads, witnesses, video—any facts or evidence that can help solve a case—are lacking. This is not surprising, explains Palmer, because "the standard practices we used to generate leads, like talking to informants or to the community or to patrol officers, haven't changed much, if at all. That kind of an approach works okay, but it relies a lot on the experience our detectives carry in their head. When the detectives retire or move on, that experience goes with them."

Palmer's conundrum was that turnover resulting from the retirement of many of his most experienced detectives was on an upward trend. True, he saw the infusion of young blood as an inherently good thing, especially given this group's increased comfort with

the new types of information—from e-mails, social media, and traffic cameras, to name a few—to which his team had access. But as Palmer recounts, the problem came when the handful of new detectives coming into the unit turned to look for guidance from the senior officers "and it's just not there. We knew at that point we needed a different way to fill the experience gap going forward."

His ad hoc efforts to come up with a solution led to blue-sky speculation. What if new detectives on the squad could pose the same questions to a computer database as they would to a veteran detective? That speculation planted a seed in Palmer's mind that wouldn't go away.

THE BIG PICTURE STARTS SMALL

What was taking shape within the robbery unit demonstrated how big ideas can come from small places. But more importantly, it showed that for these ideas to reach fruition, the "right" conditions need to be in alignment at the right time. On a leadership level, this means a driving figure in the organization who knows what it takes to nurture top-down support as well as crucial bottom-up buy-in from the ranks while keeping the department's information technology (IT) personnel on the same page. That person was Palmer. At the organizational level, the robbery unit served as a particularly good launching point for lead modeling because of the prevalence of repeat offenders among perpetrators. Ultimately, the department's ability to unleash the broader transformative potential of lead modeling would hinge in large part on the team's ability to deliver results on a small scale.

When early tests and demos proved encouraging—with the model yielding accurate results when the details of solved cases were fed into it—the team started gaining attention. The initiative received a critical boost when the robbery bureau's unit major and captain voiced their support for the direction of the project, telling Palmer that "if you can make this work, run with it." But more important than the encouragement, Palmer explains, was their willingness to advocate for the project among the department's higher-ups. "I can't get it off the ground if the brass doesn't buy in," says Palmer. "So their support was crucial."

SUCCESS BRINGS CREDIBILITY

Having been appointed the official liaison between IT and the robbery unit, Palmer set out to strengthen the case for the lead modeling tool—now officially called Blue PALMS (for Predictive Analytics Lead Modeling Software)—by building a series of successes. His constituency was not only the department brass but also the detectives whose support would be critical to its successful adoption as a robbery-solving tool. In his attempts to introduce Blue PALMS, resistance was predictably stronger among veteran detectives who saw no reason to give up their long-standing practices. Palmer knew that dictates or coercion would not win their hearts and minds. He would need to build a beachhead of credibility.

Palmer found that opportunity in one of his best and most experienced detectives. Early in a robbery investigation, the detective indicated to Palmer that he had a strong hunch who the perpetrator was and wanted, in essence, to test the Blue PALMS system. At the detective's request, the department analyst fed key details of the crime into the system, including the modus operandi (MO). The system's statistical models compared these details to a database of historical data, looking for important correlations and similarities in the crime's signature. The report that came out of the process included a list of 20 suspects ranked in order of match strength, or likelihood. When the analyst handed the detective the report, his "hunch" suspect was listed in the top five. Soon after his arrest, the suspect confessed, and Palmer had gained a solid convert.

Although it was a useful exercise, Palmer realized that the true test was not in confirming hunches but in breaking cases that had come to a dead end. Such was the situation in a carjacking that had, in Palmer's words, "no witnesses, no video, and no crime scene—nothing to go on." When the senior detective on the stalled case went on leave after three months, the junior detective to whom it was assigned requested a Blue PALMS report. Shown photographs of the top people on the suspect list, the victim made a positive identification of the suspect leading to the successful conclusion of the case. That suspect was number one on the list.

JUST THE FACTS

The success that Blue PALMS continues to build has been a major factor in Palmer's getting his detectives on board. But if there is a part of his message that resonates even more with his detectives, it is the fact that Blue PALMS is designed not to change the basics of policing practices but to enhance them by giving them a second chance of cracking the case. "Police work is at the core about human relations—about talking to witnesses, to victims, to the community—and we're not out to change that," says Palmer. "Our aim is to give investigators factual insights from information we already have that might make a difference, so even if we're successful 5 percent of the time, we're going to take a lot of offenders off the street."

The growing list of cold cases solved has helped Palmer in his efforts to reinforce the merits of Blue PALMS. But, in showing where his loyalty lies, he sees the detectives who have closed these cold cases—not the program—as most deserving of the spotlight, and that approach has gone over well. At his chief's request, Palmer is beginning to use his liaison role as a platform for reaching out to other areas in the Miami-Dade Police Department.

SAFER STREETS FOR A SMARTER CITY

When he speaks of the impact of tourism, a thread that runs through Miami-Dade's Smarter Cities vision, Palmer sees Blue PALMS as an important tool to protect one of the county's greatest assets. "The threat to tourism posed by rising street crime was a big reason the unit was established," says Palmer. "The fact that we're able to use analytics and intelligence to help us close more cases and keep more criminals off the street is good news for our citizens and our tourist industry."

➧ QUESTIONS FOR THE OPENING VIGNETTE

1. Why do law enforcement agencies and departments like the Miami-Dade Police Department embrace advanced analytics and data mining?
2. What are the top challenges for law enforcement agencies and departments like the Miami-Dade Police Department? Can you think of other challenges (not mentioned in this case) that can benefit from data mining?
3. What are the sources of data that law enforcement agencies and departments like the Miami-Dade Police Department use for their predictive modeling and data mining projects?
4. What type of analytics do law enforcement agencies and departments like the Miami-Dade Police Department use to fight crime?
5. What does "the big picture starts small" mean in this case? Explain.

WHAT WE CAN LEARN FROM THIS VIGNETTE

Law enforcement agencies and departments are under tremendous pressure to carry out their mission of safeguarding people with limited resources. The environment within which they perform their duties is becoming increasingly more challenging so that they

have to constantly adopt and perhaps stay a few steps ahead to prevent the likelihood of catastrophes. Understanding the changing nature of crime and criminals is an ongoing challenge. In the midst of these challenges, what works in favor of these agencies is the availability of the data and analytics technologies to better analyze past occurrences and to foresee future events. Data have become available more now than in the past. Applying advanced analytics and data mining tools (i.e., knowledge discovery techniques) to these large and rich data sources provides them with the insight that they need to better prepare and act on their duties. Therefore, law enforcement agencies are becoming one of the leading users of the new face of analytics. Data mining is a prime candidate for better understanding and management of these mission-critical tasks with a high level of accuracy and timeliness. The study described here clearly illustrates the power of analytics and data mining to create a holistic view of the world of crime and criminals for better and faster reaction and management. In this chapter, you will see a wide variety of data mining applications solving complex problems in a variety of industries and organizational settings where the data are used to discover actionable insight to improve mission readiness, operational efficiency, and competitive advantage.

Sources: "Miami-Dade Police Department: Predictive modeling pinpoints likely suspects based on common crime signatures of previous crimes," IBM Customer Case Studies. **www-03.ibm.com/software/businesscasestudies/om/en/corp?synkey=C894638H25952N07**; "Law Enforcement Analytics: Intelligence-Led and Predictive Policing by Information Builder." **www.informationbuilders.com/solutions/gov-lea**.

4.2 DATA MINING CONCEPTS

Data mining, a relatively new and exciting technology, has become a common practice for a vast majority of organizations. In an interview with *Computerworld* magazine in January 1999, Dr. Arno Penzias (Nobel laureate and former chief scientist of Bell Labs) identified data mining from organizational databases as a key application for corporations of the near future. In response to *Computerworld*'s age-old question of "What will be the killer applications in the corporation?" Dr. Penzias replied, "Data mining." He then added, "Data mining will become much more important and companies will throw away nothing about their customers because it will be so valuable. If you're not doing this, you're out of business." Similarly, in an article in *Harvard Business Review*, Thomas Davenport (2006) argued that the latest strategic weapon for companies is analytical decision making, providing examples of companies such as **Amazon.com**, Capital One, Marriott International, and others that have used analytics to better understand their customers and optimize their extended supply chains to maximize their returns on investment while providing the best customer service. This level of success is highly dependent on a company's thorough understanding of its customers, vendors, business processes, and the extended supply chain.

A large portion of "understanding the customer" can come from analyzing the vast amount of data that a company collects. The cost of storing and processing data has decreased dramatically in the recent past, and, as a result, the amount of data stored in electronic form has grown at an explosive rate. With the creation of large databases, the possibility of analyzing the data stored in them has emerged. The term *data mining* was originally used to describe the process through which previously unknown patterns in data were discovered. This definition has since been stretched beyond those limits by some software vendors to include most forms of data analysis in order to increase sales with the popularity of the data mining label. In this chapter, we accept the original definition of data mining.

Although the term *data mining* is relatively new, the ideas behind it are not. Many of the techniques used in data mining have their roots in traditional statistical analysis and artificial intelligence work done since the early part of the 1980s. Why, then, has it

suddenly gained the attention of the business world? Following are some of the most important reasons:

- More intense competition at the global scale driven by customers' ever-changing needs and wants in an increasingly saturated marketplace.
- General recognition of the untapped value hidden in large data sources.
- Consolidation and integration of database records, which enables a single view of customers, vendors, transactions, and so on.
- Consolidation of databases and other data repositories into a single location in the form of a data warehouse.
- The exponential increase in data processing and storage technologies.
- Significant reduction in the cost of hardware and software for data storage and processing.
- Movement toward the demassification (conversion of information resources into nonphysical form) of business practices.

Data generated by the Internet are increasing rapidly in both volume and complexity. Large amounts of genomic data are being generated and accumulated all over the world. Disciplines such as astronomy and nuclear physics create huge quantities of data on a regular basis. Medical and pharmaceutical researchers constantly generate and store data that can then be used in data mining applications to identify better ways to accurately diagnose and treat illnesses and to discover new and improved drugs.

On the commercial side, perhaps the most common use of data mining has been in the finance, retail, and healthcare sectors. Data mining is used to detect and reduce fraudulent activities, especially in insurance claims and credit card use (Chan et al., 1999); to identify customer buying patterns (Hoffman, 1999); to reclaim profitable customers (Hoffman, 1998); to identify trading rules from historical data; and to aid in increased profitability using market-basket analysis. Data mining is already widely used to better target clients, and with the widespread development of e-commerce, this can only become more imperative with time. See Application Case 4.1 for information on how Infinity P&C has used predictive analytics and data mining to improve customer service, combat fraud, and increase profit.

Application Case 4.1

Visa Is Enhancing the Customer Experience while Reducing Fraud with Predictive Analytics and Data Mining

When card issuers first started using automated business rules software to counter debit and credit card fraud, the limits on that technology were quickly evident: Customers reported frustrating payment rejections on dream vacations or critical business trips. Visa works with its clients to improve customer experience by providing cutting-edge fraud risk tools and consulting services that make its strategies more effective. Through this approach, Visa enhances customer experience and minimizes invalid transaction declines.

The company's global network connects thousands of financial institutions with millions of merchants and cardholders every day. It has been a pioneer in cashless payments for more than 50 years. By using SAS® Analytics, Visa is supporting financial institutions to reduce fraud without upsetting customers with unnecessary payment rejections. Whenever it processes a transaction, Visa analyzes up to 500 unique variables in real time to assess the risk of that transaction. Using vast data sets, including global fraud hot spots and transactional patterns, the company can more accurately assess whether you're buying escargot in Paris or someone who stole your credit card is.

"What that means is that if you are likely to travel we know it, and we tell your financial institution so you're not declined at the point of sale,"

(*Continued*)

Application Case 4.1 (Continued)

says Nathan Falkenborg, head of Visa Performance Solutions for North Asia. "We also will assist your bank in developing the right strategies for using the Visa tools and scoring systems," he adds. Visa estimates that Big Data analytics works; state-of-the-art models and scoring systems have the potential to prevent an incremental $2 billion of fraudulent payment volume annually.

A globally recognized name, Visa facilitates electronic funds transfer through branded products that are issued by its thousands of financial institution partners. The company processed 64.9 billion transactions in 2014, and $4.7 trillion in purchases were made with Visa cards in the same year.

Visa has the computing capability to process 56,000 transaction messages per second, which is more than four times the actual peak transaction rate to date. Visa does not just process and compute—it is continually using analytics to share strategic and operational insights with its partner financial institutions and assist them in improving performance. This business goal is supported by a robust data management system. Visa also assists its clients in improving performance by developing and delivering deep analytical insight.

"We understand patterns of behavior by performing clustering and segmentation at a granular level, and we provide this insight to our financial institution partners," says Falkenborg. "It's an effective way to help our clients communicate better and deepen their understanding of the customer."

As an example of marketing support, Visa has assisted clients globally in identifying segments of customers that should be offered a different Visa product. "Understanding the customer lifecycle is incredibly important, and Visa provides information to clients that help them take action and offer the right product to the right customer before a value proposition becomes stale," says Falkenborg.

How Can Using In-Memory Analytics Make a Difference?

In a recent proof of concept, Visa used a high-performance solution from SAS that relies on in-memory computing to power statistical and machine-learning algorithms and then present the information visually. In-memory analytics reduces the need to move data and perform additional model iterations, making it much faster and accurate.

Falkenborg describes the solution as like having the information memorized versus having to get up and go to a filing cabinet to retrieve it. "In-memory analytics is just taking your brain and making it bigger. Everything is instantly accessible."

Ultimately, solid analytics helps the company do more than just process payments. "We can deepen the client conversation and serve our clients even better with our incredible big data set and expertise in mining transaction data," says Falkenborg. "We use our consulting and analytics capabilities to assist our clients in tackling business challenges and protect the payment ecosystem. And that's what we do with high-performance analytics."

Falkenborg elaborates,

> The challenge that we have, as with any company managing and using massive data sets, is how we use all necessary information to solve a business challenge—whether that is improving our fraud models, or assisting a client to more effectively communicate with its customers. In-memory analytics enables us to be more nimble; with a 100× analytical system processing speed improvement, our data and decision scientists can iterate much faster.

Fast and accurate predictive analytics allows Visa to better serve clients with tailored consulting services, helping them succeed in today's fast-changing payments industry.

Questions for Case 4.1

1. What challenges were Visa and the rest of the credit card industry facing?
2. How did Visa improve customer service while also improving concepts related to retention of fraud?
3. What is in-memory analytics, and why was it necessary?

Source: "Enhancing the Customer Experience While Reducing Fraud (SAS® Analytics) High-Performance Analytics Empowers Visa to Enhance Customer Experience While Reducing Debit and Credit Card Fraud." Copyright © 2018 SAS Institute Inc., Cary, NC, USA. Reprinted with permission. All rights reserved.

Definitions, Characteristics, and Benefits

Simply defined, **data mining** is a term used to describe discovering or "mining" knowledge from large amounts of data. When considered by analogy, one can easily realize that the term *data mining* is a misnomer; that is, mining of gold from within rocks or dirt is referred to as "gold" mining rather than "rock" or "dirt" mining. Therefore, data mining perhaps should have been named "knowledge mining" or "knowledge discovery." Despite the mismatch between the term and its meaning, *data mining* has become the choice of the community. Many other names that are associated with data mining include *knowledge extraction, pattern analysis, data archaeology, information harvesting, pattern searching,* and *data dredging*.

Technically speaking, data mining is a process that uses statistical, mathematical, and artificial intelligence techniques to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data. These patterns can be in the form of business rules, affinities, correlations, trends, or prediction models (see Nemati and Barko, 2001). Most literature defines data mining as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases," where the data are organized in records structured by categorical, ordinal, and continuous variables (Fayyad, Piatetsky-Shapiro, and Smyth, 1996, pp. 40–41). In this definition, the meanings of the key term are as follows:

- *Process* implies that data mining comprises many iterative steps.
- *Nontrivial* means that some experimental type search or inference is involved; that is, it is not as straightforward as a computation of predefined quantities.
- *Valid* means that the discovered patterns should hold true on new data with a sufficient degree of certainty.
- *Novel* means that the patterns are not previously known to the user within the context of the system being analyzed.
- *Potentially useful* means that the discovered patterns should lead to some benefit to the user or task.
- *Ultimately understandable* means that the pattern should make business sense that leads to the user saying, "Mmm! It makes sense; why didn't I think of that," if not immediately, at least after some postprocessing.

Data mining is not a new discipline but rather a new definition for the use of many disciplines. Data mining is tightly positioned at the intersection of many disciplines, including statistics, artificial intelligence, machine learning, management science, information systems (IS), and databases (see Figure 4.1). Using advances in all of these disciplines, data mining strives to make progress in extracting useful information and knowledge from large databases. It is an emerging field that has attracted much attention in a very short time.

The following are the major characteristics and objectives of data mining:

- Data are often buried deep within very large databases, which sometimes contain data from several years. In many cases, the data are cleansed and consolidated into a data warehouse. Data can be presented in a variety of formats (see Chapter 3 for a brief taxonomy of data).
- The data mining environment is usually a client/server architecture or a Web-based IS architecture.
- Sophisticated new tools, including advanced visualization tools, help remove the information ore buried in corporate files or archival public records. Finding it involves massaging and synchronizing the data to get the right results. Cutting-edge data miners are also exploring the usefulness of soft data (i.e., unstructured text stored in such places as Lotus Notes databases, text files on the Internet, or enterprisewide intranets).

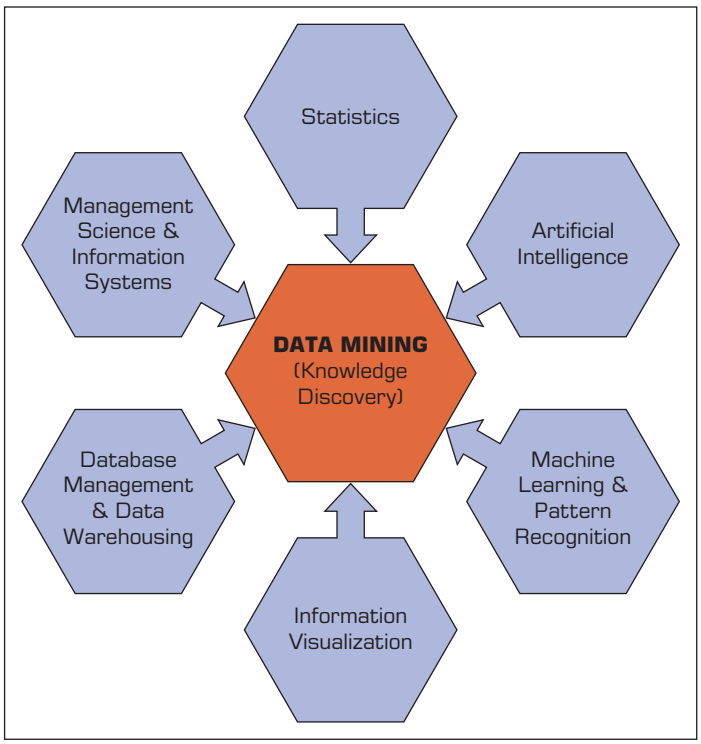


FIGURE 4.1 **Data Mining is a Blend of Multiple Disciplines.**

- The miner is often an end user empowered by data drills and other powerful query tools to ask ad hoc questions and obtain answers quickly with little or no programming skill.
- "Striking it rich" often involves finding an unexpected result and requires end users to think creatively throughout the process, including the interpretation of the findings.
- Data mining tools are readily combined with spreadsheets and other software development tools. Thus, the mined data can be analyzed and deployed quickly and easily.
- Because of the large amounts of data and massive search efforts, it is sometimes necessary to use parallel processing for data mining.

A company that effectively leverages data mining tools and technologies can acquire and maintain a strategic competitive advantage. Data mining offers organizations an indispensable decision-enhancing environment to exploit new opportunities by transforming data into a strategic weapon. See Nemati and Barko (2001) for a more detailed discussion on the strategic benefits of data mining.

How Data Mining Works

Using existing and relevant data obtained from within and outside the organization, data mining builds models to discover patterns among the attributes presented in the data set. Models are the mathematical representations (simple linear relationships/affinities and/or complex and highly nonlinear relationships) that identify the patterns among the attributes of the things (e.g., customers, events) described within the data set. Some of these patterns are explanatory (explaining the interrelationships and affinities among the attributes), whereas others are predictive (foretelling future values of certain attributes). In general, data mining seeks to identify four major types of patterns:

1. *Associations* find the commonly co-occurring groupings of things, such as beer and diapers going together in market-basket analysis.

2. *Predictions* tell the nature of future occurrences of certain events based on what has happened in the past, such as predicting the winner of the Super Bowl or forecasting the absolute temperature of a particular day.
3. *Clusters* identify natural groupings of things based on their known characteristics, such as assigning customers in different segments based on their demographics and past purchase behaviors.
4. *Sequential relationships* discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.

Application Case 4.2 shows how American Honda uses data mining (a critical component of advanced analytics tools) to enhance their understanding of the warranty claims, forecast feature part and resource needs, and better understand customer needs, wants, and opinions.

Application Case 4.2 American Honda Uses Advanced Analytics to Improve Warranty Claims

Background

When a car or truck owner brings a vehicle into an Acura or Honda dealership in the United States, there's more to the visit than a repair or a service check. During each visit, the service technicians generate data on the repairs, including any warranty claims to American Honda Motor Co., Inc., that feed directly into its database. This includes what type of work was performed, what the customer paid, service advisor comments, and many other data points.

Now, multiply this process by dozens of visits a day at over 1,200 dealerships nationwide, and it's clear—American Honda has big data. It's up to people like Kendrick Kau, assistant manager of American Honda's Advanced Analytics group, to draw insights from this data and turn it into a useful asset.

Examining Warranty Data to Make Maintenance More Efficient

Like any other major automobile distributor, American Honda works with a network of dealerships that perform warrantied repair work on its vehicles. This can be a significant cost for the company, so American Honda uses analytics to make sure that warranty claims are complete and accurate upon submission.

In the case of warranty claims, Kau's team helps empower dealers to understand the appropriate warranty processes by providing them with useful information via an online report. To support a goal of reducing inappropriate warranty costs, Kau and his team must sift through information on repairs, parts, customers, and other details. They chose a visual approach to business intelligence and analytics, powered by SAS, to identify cost reduction opportunities.

To decrease warranty expense, the Advanced Analytics team used SAS Analytics to create a proprietary process to surface suspicious warranty claims for scrutiny on a daily basis to make sure they are in compliance with existing guidelines. The effort to identify and scrutinize claims was once fairly manual, tedious, and time-intensive.

"Before SAS, it took one of our staff members one week out of each month to aggregate and report warranty data within Microsoft Excel spreadsheets," Kau says. "Now, with SAS, we populate those same reports on an easily accessible online dashboard automatically, and we recovered a week of manpower that we could put on other projects."

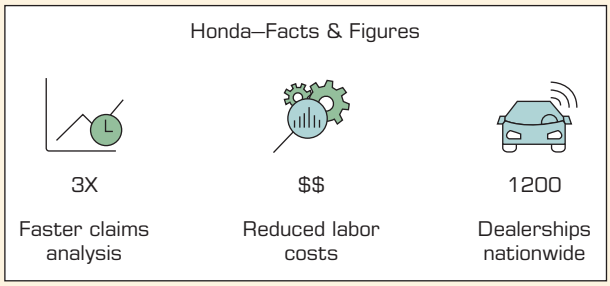
By applying SAS Analytics to warranty data, the Advanced Analytics group gave the Claims group and field personnel the ability to quickly and accurately identify claims that were incomplete, inaccurate, or noncompliant. The results were impressive.

"Initially, it took our examiners over three minutes on average to identify a potentially noncompliant claim, and even then, they were only finding a truly noncompliant claim 35 percent of the time," Kau says. "Now, with SAS, it takes less than a minute to identify a suspicious claim. And in that time, they are finding a noncompliant claim 76 percent of the time."

(*Continued*)

Application Case 4.2 (Continued)

The effort to increase warranty compliance has paid off for American Honda. Through more complete analysis of warranty claims—and more education at the dealerships—American Honda saw a reduction in labor costs for 52 percent of its available labor codes.



Using Service Data to Forecast Future Needs

The American Honda Advanced Analytics team also uses service and parts data to develop stronger bonds with customers by ensuring dealers have in-demand parts available for customer repairs. Having the right parts available—at the right time—is paramount, so vehicle repairs data feed directly into American Honda's marketing and customer retention efforts.

"For the marketing team, we provide strategic insight to help shape their programs that are designed to drive customers to the dealers, and ultimately, keep them loyal to our brand," Kau says. "The goal of Honda is lifetime owner loyalty. We want our customers to have a good experience, and one of the ways to do that is through exceptional service."

American Honda uses SAS Forecast Server to assist with business planning to ensure adequate resources are available to meet future demands for services. Using historical information on repair orders and certifications, they developed a time series using years of previous repairs. By combining time series information with sales data, Kau's team can project where the company's greatest opportunities are in the years ahead.

"Our goal is to forecast the number of vehicles in operation in order to predict the volume of customers coming into the dealerships," Kau says. "And that translates to how many parts we should have on hand and helps us to plan staffing to meet customer demands. Looking backward on a year-by-year basis, we've been within 1 percent of where we forecast to be. That's extremely good for a forecast, and I attribute much of that to the abilities of the SAS software."

Customer Feedback that Drives the Business

Another way American Honda uses analytics is to quickly evaluate customer survey data. Using SAS, the Advanced Analytics team mines survey data to gain insight into how vehicles are being used and identify design changes that are most likely to improve customer satisfaction.

On a weekly basis, the analytics team examines customer survey data. Kau's team uses SAS to flag emerging trends that may require the attention of design, manufacturing, engineering, or other groups. With SAS technology, users can drill down from high-level issues to more specific responses to understand a potential root cause.

"We can look into the data and see what the customers are saying," Kau says. "And that leads to a number of questions that we can tackle. Is a component designed in the most optimal way? Is it a customer education issue? Is it something that we should address at the manufacturing process? Because of SAS, these are critical questions that we can now identify using our data."

Questions for Case 4.2

1. How does American Honda use analytics to improve warranty claims?
2. In addition to warranty claims, for what other purposes does American Honda use advanced analytics methods?
3. Can you think of other uses of advanced analytics in the automotive industry? You can search the Web to find some answers to this question.

Source: SAS Case Study "American Honda Motor Co., Inc. uses SAS advanced analytics to improve warranty claims" **https://www.sas.com/en_us/customers/american-honda.html** (accessed June 2018)

These types of patterns have been *manually* extracted from data by humans for centuries, but the increasing volume of data in modern times has created the need for more automatic approaches. As data sets have grown in size and complexity, direct manual data analysis has increasingly been augmented with indirect, automatic data-processing tools that use sophisticated methodologies, methods, and algorithms. The manifestation of such evolution of automated and semiautomated means of processing large data sets is now commonly referred to as *data mining*.

Generally speaking, data mining tasks can be classified into three main categories: prediction, association, and clustering. Based on the way in which the patterns are extracted from the historical data, the learning algorithms of data mining methods can be classified as either supervised or unsupervised. With supervised learning algorithms, the training data includes both the descriptive attribute (i.e., independent variables or decision variables) and the class attribute (i.e., output variable or result variable). In contrast, with unsupervised learning, the training includes only descriptive attributes. Figure 4.2 shows a simple taxonomy for data mining tasks along with the learning methods and popular algorithms for each of the data mining tasks.

PREDICTION **Prediction** is commonly referred to as the act of telling about the future. It differs from simple guessing by taking into account the experiences, opinions, and other relevant information in conducting the task of foretelling. A term that is commonly associated with prediction is *forecasting*. Even though many believe that these two terms are synonymous, there is a subtle but critical difference between the two. Whereas prediction is largely experience and opinion based, forecasting is data and model based. That is, in order of increasing reliability, one might list the relevant terms as *guessing, predicting,* and *forecasting*, respectively. In data mining terminology, *prediction* and *forecasting* are used synonymously, and the term *prediction* is used as the common representation of the act. Depending on the nature of what is being predicted, prediction can be named more specifically as classification (where the predicted thing, such as tomorrow's forecast, is a class label such as "rainy" or "sunny") or regression (where the predicted thing, such as tomorrow's temperature, is a real number, such as "65°F").

CLASSIFICATION **Classification**, or supervised induction, is perhaps the most common of all data mining tasks. The objective of classification is to analyze historical data stored in a database and automatically generate a model that can predict future behavior. This induced model consists of generalizations over the records of a training data set, which help distinguish predefined classes. The hope is that the model can then be used to predict the classes of other unclassified records and, more importantly, to accurately predict actual future events.

Common classification tools include neural networks and decision trees (from machine learning), logistic regression and discriminant analysis (from traditional statistics), and emerging tools such as rough sets, support vector machines (SVMs), and genetic algorithms. Statistics-based classification techniques (e.g., logistic regression and discriminant analysis) have received their share of criticism—that they make unrealistic assumptions about the data, such as independence and normality—which limit their use in classification-type data mining projects.

Neural networks involve the development of mathematical structures (somewhat resembling the biological neural networks in the human brain) that have the capability to learn from past experiences presented in the form of well-structured data sets. They tend to be more effective when the number of variables involved is rather large and the relationships among them are complex and imprecise. Neural networks have disadvantages as well as advantages. For example, providing a good rationale for the predictions made by a neural network is usually very difficult. Also, training neural networks usually

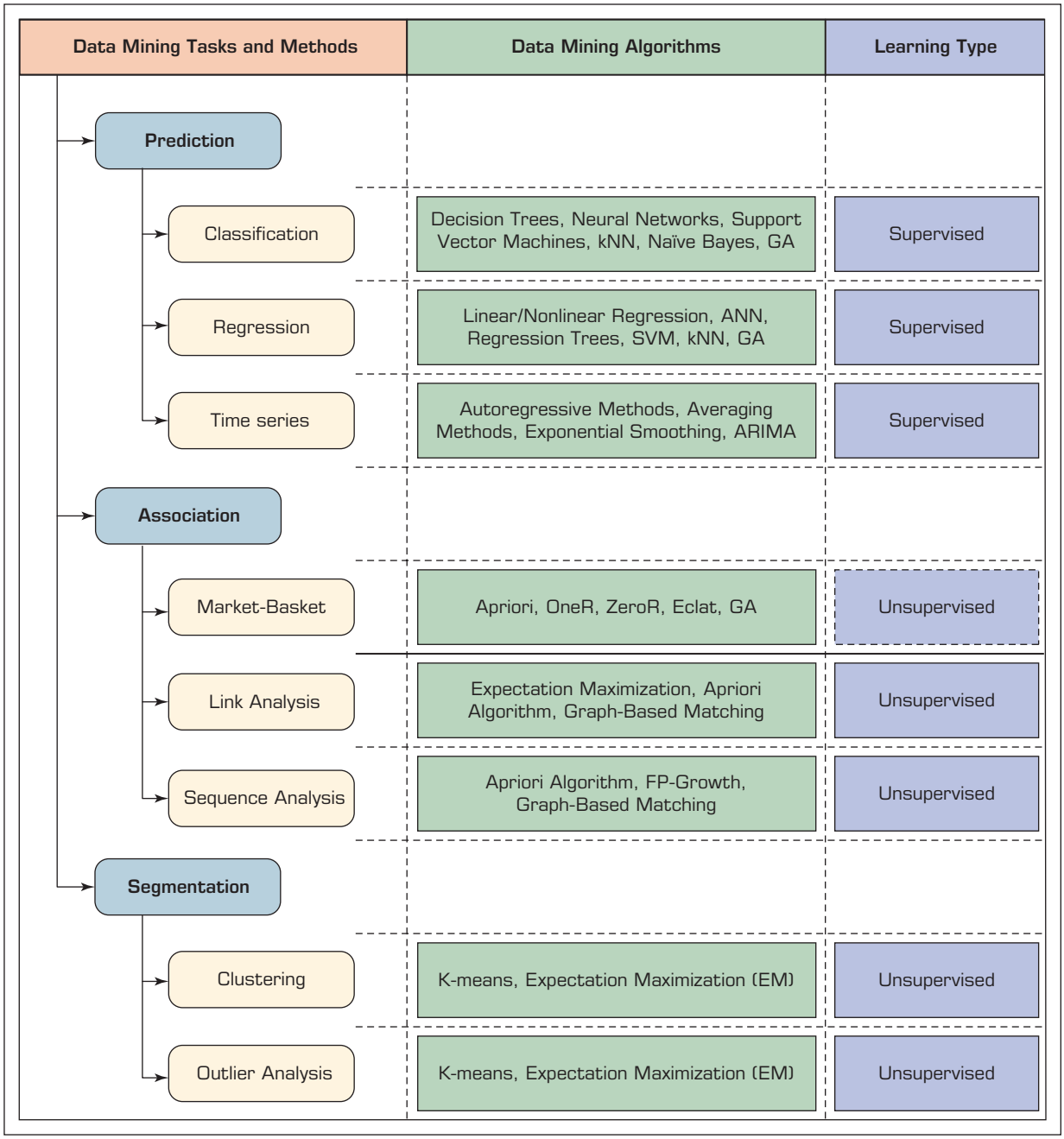


FIGURE 4.2 **Simple Taxonomy for Data Mining Tasks, Methods, and Algorithms.**

takes a considerable amount of time. Unfortunately, the time needed for training tends to increase exponentially as the volume of data increases, and in general, neural networks cannot be trained on very large databases. These and other factors have limited the applicability of neural networks in data-rich domains.

Decision trees classify data into a finite number of classes based on the values of the input variables. Decision trees are essentially a hierarchy of if-then statements and are thus significantly faster than neural networks. They are most appropriate for **categorical data** and **interval data**. Therefore, incorporating continuous variables into a decision

tree framework requires *discretization*, that is, converting continuous valued numerical variables to ranges and categories.

A related category of classification tools is rule induction. Unlike with a decision tree, with rule induction the if-then statements are induced from the training data directly, and they need not be hierarchical in nature. Other, more recent techniques such as SVM, rough sets, and genetic algorithms are gradually finding their way into the arsenal of classification algorithms.

CLUSTERING **Clustering** partitions a collection of things (e.g., objects, events, presented in a structured data set) into segments (or natural groupings) whose members share similar characteristics. Unlike in classification, in clustering, the class labels are unknown. As the selected algorithm goes through the data set, identifying the commonalities of things based on their characteristics, the clusters are established. Because the clusters are determined using a heuristic-type algorithm and because different algorithms could end up with different sets of clusters for the same data set, before the results of clustering techniques are put to actual use, it could be necessary for an expert to interpret, and potentially modify, the suggested clusters. After reasonable clusters have been identified, they can be used to classify and interpret new data.

Not surprisingly, clustering techniques include optimization. The goal of clustering is to create groups so that the members within each group have maximum similarity and the members across groups have minimum similarity. The most commonly used clustering techniques include *k*-means (from statistics) and self-organizing maps (from machine learning), which is a unique neural network architecture developed by Kohonen (1982).

Firms often effectively use their data mining systems to perform market segmentation with cluster analysis. Cluster analysis is a means of identifying classes of items so that items in a cluster have more in common with each other than with items in other clusters. Cluster analysis can be used in segmenting customers and directing appropriate marketing products to the segments at the right time in the right format at the right price. Cluster analysis is also used to identify natural groupings of events or objects so that a common set of characteristics of these groups can be identified to describe them.

ASSOCIATIONS **Associations**, or *association rule learning in data mining*, is a popular and well-researched technique for discovering interesting relationships among variables in large databases. Thanks to automated data-gathering technologies such as bar code scanners, the use of association rules for discovering regularities among products in large-scale transactions recorded by point-of-sale systems in supermarkets has become a common knowledge discovery task in the retail industry. In the context of the retail industry, association rule mining is often called *market-basket analysis*.

Two commonly used derivatives of association rule mining are **link analysis** and **sequence mining**. With link analysis, the linkage among many objects of interest is discovered automatically, such as the link between Web pages and referential relationships among groups of academic publication authors. With sequence mining, relationships are examined in terms of their order of occurrence to identify associations over time. Algorithms used in association rule mining include the popular Apriori (where frequent itemsets are identified) and FP-Growth, OneR, ZeroR, and Eclat.

VISUALIZATION AND TIME-SERIES FORECASTING Two techniques often associated with data mining are *visualization* and *time-series forecasting*. Visualization can be used in conjunction with other data mining techniques to gain a clearer understanding of underlying relationships. As the importance of visualization has increased in recent years, a new term, *visual analytics*, has emerged. The idea is to combine analytics and visualization in a single environment for easier and faster knowledge creation. Visual analytics is covered in

detail in Chapter 3. In time-series forecasting, the data consist of values of the same variable that are captured and stored over time in regular intervals. These data are then used to develop forecasting models to extrapolate the future values of the same variable.

Data Mining Versus Statistics

Data mining and statistics have a lot in common. They both look for relationships within data. Most people call statistics the "foundation of data mining." The main difference between the two is that statistics starts with a well-defined proposition and hypothesis whereas data mining starts with a loosely defined discovery statement. Statistics collects sample data (i.e., primary data) to test the hypothesis whereas data mining and analytics use all the existing data (i.e., often observational, secondary data) to discover novel patterns and relationships. Another difference comes from the size of data that they use. Data mining looks for data sets that are as "big" as possible, whereas statistics looks for the right size of data (if the data are larger than what is needed/required for the statistical analysis, a sample of them is used). The meaning of "large data" is rather different between statistics and data mining. A few hundred to a thousand data points are large enough to a statistician, but several million to a few billion data points are considered large for data mining studies.

SECTION 4.2 REVIEW QUESTIONS

1. Define *data mining*. Why are there many different names and definitions for data mining?
2. What recent factors have increased the popularity of data mining?
3. Is data mining a new discipline? Explain.
4. What are some major data mining methods and algorithms?
5. What are the key differences between the major data mining tasks?

4.3 DATA MINING APPLICATIONS

Data mining has become a popular tool in addressing many complex business problems and opportunities. It has been proven to be very successful and helpful in many areas, some of which are shown by the following representative examples. The goal of many of these business data mining applications is to solve a pressing problem or to explore an emerging business opportunity to create a sustainable competitive advantage.

- **Customer relationship management.** CRM is the extension of traditional marketing. The goal of CRM is to create one-on-one relationships with customers by developing an intimate understanding of their needs and wants. As businesses build relationships with their customers over time through a variety of interactions (e.g., product inquiries, sales, service requests, warranty calls, product reviews, social media connections), they accumulate tremendous amounts of data. When combined with demographic and socioeconomic attributes, this information-rich data can be used to (1) identify most likely responders/buyers of new products/services (i.e., customer profiling), (2) understand the root causes of customer attrition to improve customer retention (i.e., churn analysis), (3) discover time-variant associations between products and services to maximize sales and customer value, and (4) identify the most profitable customers and their preferential needs to strengthen relationships and to maximize sales.
- **Banking.** Data mining can help banks with the following: (1) automating the loan application process by accurately predicting the most probable defaulters,

(2) detecting fraudulent credit card and online banking transactions, (3) identifying ways to maximize value for customers by selling them products and services that they are most likely to buy, and (4) optimizing the cash return by accurately forecasting the cash flow on banking entities (e.g., ATM machines, banking branches).

- **Retailing and logistics.** In the retailing industry, data mining can be used to (1) predict accurate sales volumes at specific retail locations to determine correct inventory levels, (2) identify sales relationships between different products (with market-basket analysis) to improve the store layout and optimize sales promotions, (3) forecast consumption levels of different product types (based on seasonal and environmental conditions) to optimize logistics and, hence, maximize sales, and (4) discover interesting patterns in the movement of products (especially for products that have a limited shelf life because they are prone to expiration, perishability, and contamination) in a supply chain by analyzing sensory and radio-frequency identification (RFID) data.
- **Manufacturing and production.** Manufacturers can use data mining to (1) predict machinery failures before they occur through the use of sensory data (enabling what is called *condition-based maintenance*), (2) identify anomalies and commonalities in production systems to optimize manufacturing capacity, and (3) discover novel patterns to identify and improve product quality.
- **Brokerage and securities trading.** Brokers and traders use data mining to (1) predict when and how much certain bond prices will change, (2) forecast the range and direction of stock fluctuations, (3) assess the effect of particular issues and events on overall market movements, and (4) identify and prevent fraudulent activities in securities trading.
- **Insurance.** The insurance industry uses data mining techniques to (1) forecast claim amounts for property and medical coverage costs for better business planning, (2) determine optimal rate plans based on the analysis of claims and customer data, (3) predict which customers are more likely to buy new policies with special features, and (4) identify and prevent incorrect claim payments and fraudulent activities.
- **Computer hardware and software.** Data mining can be used to (1) predict disk drive failures well before they actually occur, (2) identify and filter unwanted Web content and e-mail messages, (3) detect and prevent computer network security breaches, and (4) identify potentially unsecure software products.
- **Government and defense.** Data mining also has a number of military applications. It can be used to (1) forecast the cost of moving military personnel and equipment, (2) predict an adversary's moves and, hence, develop more successful strategies for military engagements, (3) predict resource consumption for better planning and budgeting, and (4) identify classes of unique experiences, strategies, and lessons learned from military operations for better knowledge sharing throughout the organization.
- **Travel industry (airlines, hotels/resorts, rental car companies).** Data mining has a variety of uses in the travel industry. It is successfully used to (1) predict sales of different services (seat types in airplanes, room types in hotels/resorts, car types in rental car companies) in order to optimally price services to maximize revenues as a function of time-varying transactions (commonly referred to as *yield management*), (2) forecast demand at different locations to better allocate limited organizational resources, (3) identify the most profitable customers and provide them with personalized services to maintain their repeat business, and (4) retain valuable employees by identifying and acting on the root causes for attrition.
- **Healthcare.** Data mining has a number of healthcare applications. It can be used to (1) identify people without health insurance and the factors underlying this undesired phenomenon, (2) identify novel cost–benefit relationships between different

treatments to develop more effective strategies, (3) forecast the level and the time of demand at different service locations to optimally allocate organizational resources, and (4) understand the underlying reasons for customer and employee attrition.

- **Medicine.** Use of data mining in medicine should be viewed as an invaluable complement to traditional medical research, which is mainly clinical and biological in nature. Data mining analyses can (1) identify novel patterns to improve survivability of patients with cancer, (2) predict success rates of organ transplantation patients to develop better organ donor matching policies, (3) identify the functions of different genes in the human chromosome (known as *genomics*), and (4) discover the relationships between symptoms and illnesses (as well as illnesses and successful treatments) to help medical professionals make informed and correct decisions in a timely manner.
- **Entertainment industry.** Data mining is successfully used by the entertainment industry to (1) analyze viewer data to decide what programs to show during prime time and how to maximize returns by knowing where to insert advertisements, (2) predict the financial success of movies before they are produced to make investment decisions and to optimize the returns, (3) forecast the demand at different locations and different times to better schedule entertainment events and to optimally allocate resources, and (4) develop optimal pricing policies to maximize revenues.
- **Homeland security and law enforcement.** Data mining has a number of homeland security and law enforcement applications. It is often used to (1) identify patterns of terrorist behaviors (see Application Case 4.3 for an example of the use of data mining to track funding of terrorists' activities), (2) discover crime patterns (e.g., locations, timings, criminal behaviors, and other related attributes) to help solve criminal cases in a timely manner, (3) predict and eliminate potential biological and chemical attacks to the nation's critical infrastructure by analyzing special-purpose sensory data, and (4) identify and stop malicious attacks on critical information infrastructures (often called *information warfare*).

Application Case 4.3

Predictive Analytic and Data Mining Help Stop Terrorist Funding

The terrorist attack on the World Trade Center on September 11, 2001, underlined the importance of open source intelligence. The USA PATRIOT Act and the creation of the U.S. Department of Homeland Security heralded the potential application of information technology and data mining techniques to detect money laundering and other forms of terrorist financing. Law enforcement agencies had been focusing on money laundering activities via normal transactions through banks and other financial service organizations.

Law enforcement agencies are now focusing on international trade pricing as a terrorism funding tool. Money launderers have used international trade to move money silently out of a country without attracting government attention. They achieve this transfer by overvaluing imports and undervaluing exports. For example, a domestic importer and foreign exporter could form a partnership and overvalue imports, thereby transferring money from the home country, resulting in crimes related to customs fraud, income tax evasion, and money laundering. The foreign exporter could be a member of a terrorist organization.

Data mining techniques focus on analysis of data on import and export transactions from the U.S. Department of Commerce and commerce-related entities. Import prices that exceed the upper quartile of import prices and export prices that are lower than the lower quartile of export prices are tracked. The focus is on abnormal transfer prices between corporations that might result in shifting

taxable income and taxes out of the United States. An observed price deviation could be related to income tax avoidance/evasion, money laundering, or terrorist financing. The observed price deviation could also be due to an error in the U.S. trade database.

Data mining will result in efficient evaluation of data, which, in turn, will aid in the fight against terrorism. The application of information technology and data mining techniques to financial transactions can contribute to better intelligence information.

Questions for Case 4.3

1. How can data mining be used to fight terrorism? Comment on what else can be done beyond what is covered in this short application case.
2. Do you think data mining, although essential for fighting terrorist cells, also jeopardizes individuals' rights of privacy?

Sources: J. S. Zdanowic, "Detecting Money Laundering and Terrorist Financing via Data Mining," *Communications of the ACM, 47*(5), May 2004, p. 53; R. J. Bolton, "Statistical Fraud Detection: A Review," *Statistical Science, 17*(3), January 2002, p. 235.

- **Sports.** Data mining was used to improve the performance of National Basketball Association (NBA) teams in the United States. Major League Baseball teams are into predictive analytics and data mining to optimally utilize their limited resources for a winning season. In fact, most, if not all, professional sports today employ data crunchers and use data mining to increase their chances of winning. Data mining applications are not limited to professional sports. In an article, Delen at al. (2012) developed data mining models to predict National Collegiate Athletic Association (NCAA) Bowl Game outcomes using a wide range of variables about the two opposing teams' previous games statistics (more details about this case study are provided in Chapter 3). Wright (2012) used a variety of predictors for examination of the NCAA men's basketball championship (a.k.a. March Madness) bracket.

SECTION 4.3 REVIEW QUESTIONS

1. What are the major application areas for data mining?
2. Identify at least five specific applications of data mining and list five common characteristics of these applications.
3. What do you think is the most prominent application area for data mining? Why?
4. Can you think of other application areas for data mining not discussed in this section? Explain.

4.4 DATA MINING PROCESS

To systematically carry out data mining projects, a general process is usually followed. Based on best practices, data mining researchers and practitioners have proposed several processes (workflows or simple step-by-step approaches) to maximize the chances of success in conducting data mining projects. These efforts have led to several standardized processes, some of which (a few of the most popular ones) are described in this section.

One such standardized process, arguably the most popular one, the Cross-Industry Standard Process for Data Mining—**CRISP-DM**—was proposed in the mid-1990s by a European consortium of companies to serve as a nonproprietary standard methodology for data mining (CRISP-DM, 2013). Figure 4.3 illustrates this proposed process, which is a sequence of six steps that starts with a good understanding of the business and the

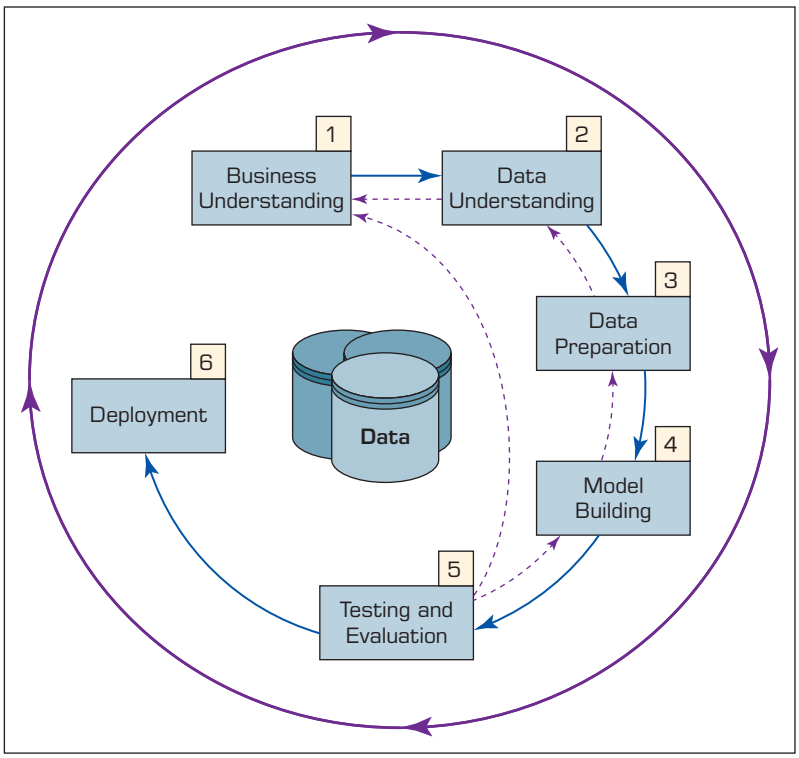


FIGURE 4.3 **Six-Step CRISP-DM Data Mining Process.**

need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfies the specific business need. Even though these steps are sequential in nature, there is usually a great deal of backtracking. Because data mining is driven by experience and experimentation, depending on the problem situation and the knowledge/experience of the analyst, the whole process can be very iterative (i.e., one should expect to go back and forth through the steps quite a few times) and time consuming. Because later steps are built on the outcomes of the former ones, one should pay extra attention to the earlier steps in order to not put the whole study on an incorrect path from the onset.

Step 1: Business Understanding

The key element of any data mining study is to know what the study is for. Determining this begins with a thorough understanding of the managerial need for new knowledge and an explicit specification of the business objective regarding the study to be conducted. Specific goals answering questions such as "What are the common characteristics of the customers we have lost to our competitors recently?" or "What are typical profiles of our customers, and how much value does each of them provide to us?" are needed. Then a project plan for finding such knowledge is developed that specifies the people responsible for collecting the data, analyzing the data, and reporting the findings. At this early stage, a budget to support the study should also be established at least at a high level with rough numbers.

Step 2: Data Understanding

A data mining study is specific to addressing a well-defined business task, and different business tasks require different sets of data. Following the business understanding

step, the main activity of the data mining process is to identify the relevant data from many available databases. Some key points must be considered in the data identification and selection phase. First and foremost, the analyst should be clear and concise about the description of the data mining task so that the most relevant data can be identified. For example, a retail data mining project could seek to identify spending behaviors of female shoppers who purchase seasonal clothes based on their demographics, credit card transactions, and socioeconomic attributes. Furthermore, the analyst should build an intimate understanding of the data sources (e.g., where the relevant data are stored and in what form; what the process of collecting the data is—automated versus manual; who the collectors of the data are and how often the data are updated) and the variables (e.g., What are the most relevant variables? Are there any synonymous and/or homonymous variables? Are the variables independent of each other—do they stand as a complete information source without overlapping or conflicting information?).

To better understand the data, the analyst often uses a variety of statistical and graphical techniques, such as simple statistical summaries of each variable (e.g., for numeric variables, the average, minimum/maximum, median, and standard deviation are among the calculated measures whereas for categorical variables, the mode and frequency tables are calculated), and correlation analysis, scatterplots, histograms, and box plots can be used. A careful identification and selection of data sources and the most relevant variables can make it easier for data mining algorithms to quickly discover useful knowledge patterns.

Data sources for data selection can vary. Traditionally, data sources for business applications include demographic data (such as income, education, number of households, and age), sociographic data (such as hobby, club membership, and entertainment), transactional data (sales record, credit card spending, issued checks), and so on. Today, data sources also use external (open or commercial) data repositories, social media, and machine-generated data.

Data can be categorized as quantitative and qualitative. Quantitative data are measured using numeric values, or **numeric data**. They can be discrete (such as integers) or continuous (such as real numbers). Qualitative data, also known as categorical data, contain both nominal and ordinal data. **Nominal data** have finite nonordered values (e.g., gender data, which have two values: male and female). **Ordinal data** have finite ordered values. For example, customer credit ratings are considered ordinal data because the ratings can be excellent, fair, and bad. A simple taxonomy of data (i.e., the nature of data) is provided in Chapter 3.

Quantitative data can be readily represented by some sort of probability distribution. A probability distribution describes how the data are dispersed and shaped. For instance, normally distributed data are symmetric and are commonly referred to as being a bell-shaped curve. Qualitative data can be coded to numbers and then described by frequency distributions. Once the relevant data are selected according to the data mining business objective, data preprocessing should be pursued.

Step 3: Data Preparation

The purpose of data preparation (more commonly called *data preprocessing*) is to take the data identified in the previous step and prepare it for analysis by data mining methods. Compared to the other steps in CRISP-DM, data preprocessing consumes the most time and effort; most people believe that this step accounts for roughly 80 percent of the total time spent on a data mining project. The reason for such an enormous effort spent on this step is the fact that real-world data are generally incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data),

noisy (containing errors or outliers), and inconsistent (containing discrepancies in codes or names). The nature of the data and the issues related to the preprocessing of data for analytics are explained in detail in Chapter 3.

Step 4: Model Building

In this step, various modeling techniques are selected and applied to an already prepared data set to address the specific business need. The model-building step also encompasses the assessment and comparative analysis of the various models built. Because there is not a universally known *best* method or algorithm for a data mining task, one should use a variety of viable model types along with a well-defined experimentation and assessment strategy to identify the "best" method for a given purpose. Even for a single method or algorithm, a number of parameters need to be calibrated to obtain optimal results. Some methods could have specific requirements in the way that the data are to be formatted; thus, stepping back to the data preparation step is often necessary. Application Case 4.4 presents a research study in which a number of model types are developed and compared to each other.

Application Case 4.4 Data Mining Helps in Cancer Research

According to the American Cancer Society, half of all men and one-third of all women in the United States will develop cancer during their lifetimes; approximately 1.5 million new cancer cases were expected to be diagnosed in 2013. Cancer is the second most common cause of death in the United States and in the world, exceeded only by cardiovascular disease. This year, more than 500,000 Americans are expected to die of cancer—more than 1,300 people a day—accounting for nearly one of every four deaths.

Cancer is a group of diseases generally characterized by uncontrolled growth and spread of abnormal cells. If the growth and/or spread are not controlled, cancer can result in death. Even though the exact reasons are not known, cancer is believed to be caused by both external factors (e.g., tobacco, infectious organisms, chemicals, and radiation) and internal factors (e.g., inherited mutations, hormones, immune conditions, and mutations that occur from metabolism). These causal factors can act together or in sequence to initiate or promote carcinogenesis. Cancer is treated with surgery, radiation, chemotherapy, hormone therapy, biological therapy, and targeted therapy. Survival statistics vary greatly by cancer type and stage at diagnosis.

The five-year relative survival rate for all cancers is improving, and the decline in cancer mortality had reached 20 percent in 2013, translating into the avoidance of about 1.2 million deaths from cancer since 1991. That's more than 400 lives saved per day! The improvement in survival reflects progress in diagnosing certain cancers at an earlier stage and improvements in treatment. Further improvements are needed to prevent and treat cancer.

Even though cancer research has traditionally been clinical and biological in nature, in recent years, data-driven analytic studies have become a common complement. In medical domains where data- and analytics-driven research has been applied successfully, novel research directions have been identified to further advance the clinical and biological studies. Using various types of data, including molecular, clinical, literature-based, and clinical trial data, along with suitable data mining tools and techniques, researchers have been able to identify novel patterns, paving the road toward a cancer-free society.

In one study, Delen (2009) used three popular data mining techniques (decision trees, artificial neural networks, and SVMs) in conjunction with logistic regression to develop prediction models for prostate cancer survivability. The data set contained around

120,000 records and 77 variables. A *k*-fold cross-validation methodology was used in model building, evaluation, and comparison. The results showed that support vector models are the most accurate predictor (with a test set accuracy of 92.85%) for this domain followed by artificial neural networks and decision trees. Furthermore, using a sensitivity–analysis-based evaluation method, the study also revealed novel patterns related to prognostic factors of prostate cancer.

In a related study, Delen, Walker, and Kadam (2005) used two data mining algorithms (artificial neural networks and decision trees) and logistic regression to develop prediction models for breast cancer survival using a large data set (more than 200,000 cases). Using a 10-fold cross-validation method to measure the unbiased estimate of the prediction models for performance comparison purposes, the researchers determined that the results indicated that the decision tree (C5 algorithm) was the best predictor with 93.6 percent accuracy on the holdout sample (which was the best prediction accuracy reported in the literature) followed by artificial neural networks with 91.2 percent accuracy, and logistic regression, with 89.2 percent accuracy. Further analysis of prediction models revealed prioritized importance of the prognostic factors, which can then be used as a basis for further clinical and biological research studies.

In the most recent study, Zolbanin et al. (2015) studied the impact of comorbidity in cancer survivability. Although prior research has shown that diagnostic and treatment recommendations might be altered based on the severity of comorbidities, chronic diseases are still being investigated in isolation from one another in most cases. To illustrate the significance of concurrent chronic diseases in the course of treatment, their study used the Surveillance, Epidemiology, and End Results (SEER) Program's cancer data to create two comorbid data sets: one for breast and female genital cancers and another for prostate and urinal cancers. Several popular machine-learning techniques are then applied to the resultant data sets to build predictive models (see Figure 4.4). Comparison of the results has shown that having more information about comorbid conditions of patients can improve models' predictive power, which in turn can help practitioners make better diagnostic and treatment decisions. Therefore, the study suggested that proper identification, recording, and use of patients' comorbidity status can potentially lower treatment costs and ease the healthcare-related economic challenges.

These examples (among many others in the medical literature) show that advanced data mining techniques can be used to develop models that possess a high degree of predictive as well as explanatory power. Although data mining methods are capable of extracting patterns and relationships hidden deep in large and complex medical databases, without the cooperation and feedback from medical experts, their results are not of much use. The patterns found via data mining methods should be evaluated by medical professionals who have years of experience in the problem domain to decide whether they are logical, actionable, and novel enough to warrant new research directions. In short, data mining is not meant to replace medical professionals and researchers but to complement their invaluable efforts to provide data-driven new research directions and to ultimately save more human lives.

Questions for Case 4.4

1. How can data mining be used for ultimately curing illnesses like cancer?
2. What do you think are the promises and major challenges for data miners in contributing to medical and biological research endeavors?

Sources: H. M. Zolbanin, D. Delen, & A. H. Zadeh, "Predicting Overall Survivability in Comorbidity of Cancers: A Data Mining Approach," *Decision Support Systems, 74,* 2015, pp. 150–161; D. Delen, "Analysis of Cancer Data: A Data Mining Approach," *Expert Systems, 26*(1), 2009, pp. 100–112; J. Thongkam, G. Xu, Y. Zhang, & F. Huang, "Toward Breast Cancer Survivability Prediction Models Through Improving Training Space," *Expert Systems with Applications, 36*(10), 2009, pp. 12200–12209; D. Delen, G. Walker, & A. Kadam, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," *Artificial Intelligence in Medicine, 34*(2), 2005, pp. 113–127.

(*Continued*)

Application Case 4.4 (Continued)

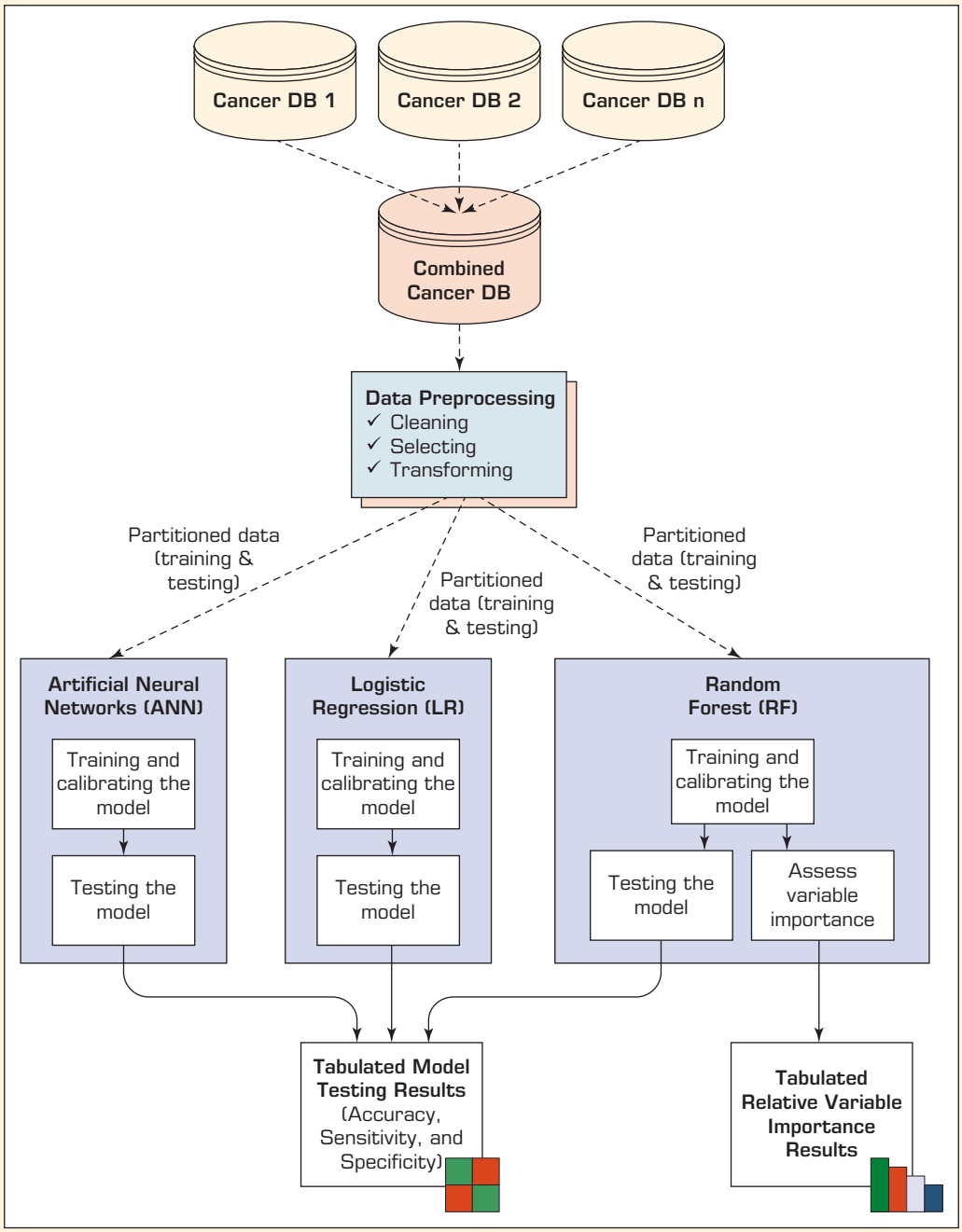


FIGURE 4.4 Data Mining Methodology for Investigation of Comorbidity in Cancer Survivability.

Depending on the business need, the data mining task can be of a prediction (either classification or regression), an association, or a clustering type. Each of these data mining tasks can use a variety of data mining methods and algorithms. Some of these data mining methods were explained earlier in this chapter, and some of the most popular

algorithms, including decision trees for classification, *k*-means for clustering, and the Apriori algorithm for association rule mining, are described later in this chapter.

Step 5: Testing and Evaluation

In step 5, the developed models are assessed and evaluated for their accuracy and generality. This step assesses the degree to which the selected model (or models) meets the business objectives and, if so, to what extent (i.e., Do more models need to be developed and assessed?). Another option is to test the developed model(s) in a real-world scenario if time and budget constraints permit. Even though the outcome of the developed models is expected to relate to the original business objectives, other findings that are not necessarily related to the original business objectives but that might also unveil additional information or hints for future directions often are discovered.

The testing and evaluation step is a critical and challenging task. No value is added by the data mining task until the business value obtained from discovered knowledge patterns is identified and recognized. Determining the business value from discovered knowledge patterns is somewhat similar to playing with puzzles. The extracted knowledge patterns are pieces of the puzzle that need to be put together in the context of the specific business purpose. The success of this identification operation depends on the interaction among data analysts, business analysts, and decision makers (such as business managers). Because data analysts might not have the full understanding of the data mining objectives and what they mean to the business and the business analysts, and decision makers might not have the technical knowledge to interpret the results of sophisticated mathematical solutions, interaction among them is necessary. To properly interpret knowledge patterns, it is often necessary to use a variety of tabulation and visualization techniques (e.g., pivot tables, cross-tabulation of findings, pie charts, histograms, box plots, scatterplots).

Step 6: Deployment

Development and assessment of the models is not the end of the data mining project. Even if the purpose of the model is to have a simple exploration of the data, the knowledge gained from such exploration will need to be organized and presented in a way that the end user can understand and benefit from. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will not carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out to actually make use of the created models.

The deployment step can also include maintenance activities for the deployed models. Because everything about the business is constantly changing, the data that reflect the business activities also are changing. Over time, the models (and the patterns embedded within them) built on the old data can become obsolete, irrelevant, or misleading. Therefore, monitoring and maintenance of the models are important if the data mining results are to become a part of the day-to-day business and its environment. A careful preparation of a maintenance strategy helps avoid unnecessarily long periods of incorrect usage of data mining results. To monitor the deployment of the data mining result(s), the project needs a detailed plan on the monitoring process, which might not be a trivial task for complex data mining models.

Other Data Mining Standardized Processes and Methodologies

To be applied successfully, a data mining study must be viewed as a process that follows a standardized methodology rather than as a set of automated software tools and techniques. In addition to CRISP-DM, there is another well-known methodology developed

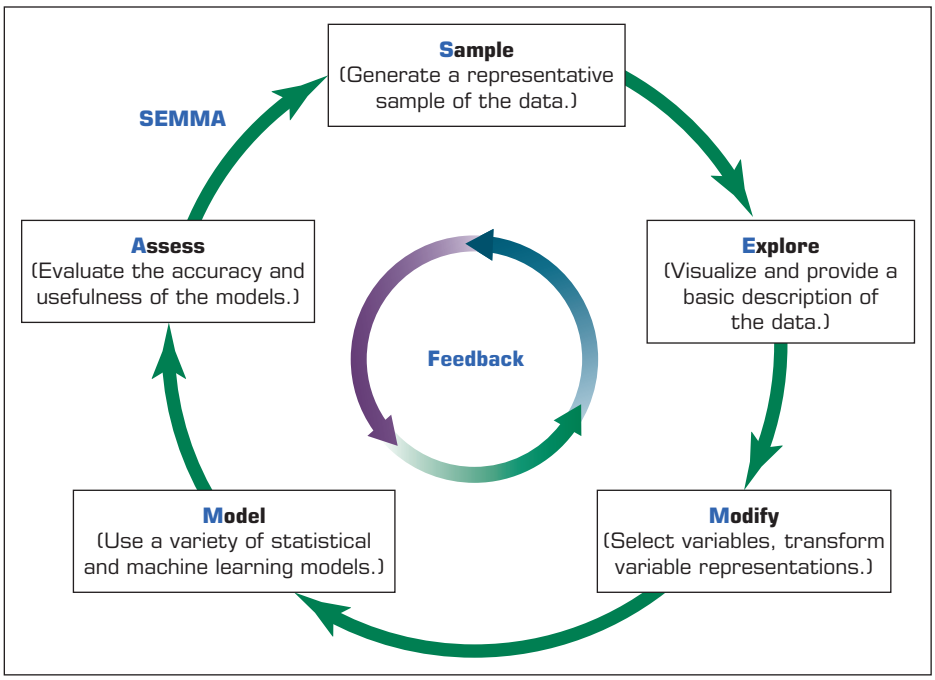


FIGURE 4.5 **SEMMA Data Mining Process.**

by the SAS Institute, called SEMMA (2009). The acronym **SEMMA** stands for "sample, explore, modify, model, and assess."

Beginning with a statistically representative sample of the data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy. A pictorial representation of SEMMA is given in Figure 4.5.

By assessing the outcome of each stage in the SEMMA process, the model developer can determine how to model new questions raised by the previous results and thus proceed back to the exploration phase for additional refinement of the data; that is, as with CRISP-DM, SEMMA is driven by a highly iterative experimentation cycle. The main difference between CRISP-DM and SEMMA is that CRISP-DM takes a more comprehensive approach—including understanding of the business and the relevant data—to data mining projects whereas SEMMA implicitly assumes that the data mining project's goals and objectives along with the appropriate data sources have been identified and understood.

Some practitioners commonly use the term **knowledge discovery in databases (KDD)** as a synonym for data mining. Fayyad et al. (1996) defined *knowledge discovery in databases* as a process of using data mining methods to find useful information and patterns in the data as opposed to data mining, which involves using algorithms to identify patterns in data derived through the KDD process (see Figure 4.6). KDD is a comprehensive process that encompasses data mining. The input to the KDD process consists of organizational data. The enterprise data warehouse enables KDD to be implemented efficiently because it provides a single source for data to be mined. Dunham (2003) summarized the KDD process as consisting of the following steps: data selection, data preprocessing, data transformation, data mining, and interpretation/evaluation.

Figure 4.7 shows the polling results for the question, "What main methodology are you using for data mining?" (conducted by **KDnuggets.com** in August 2007).

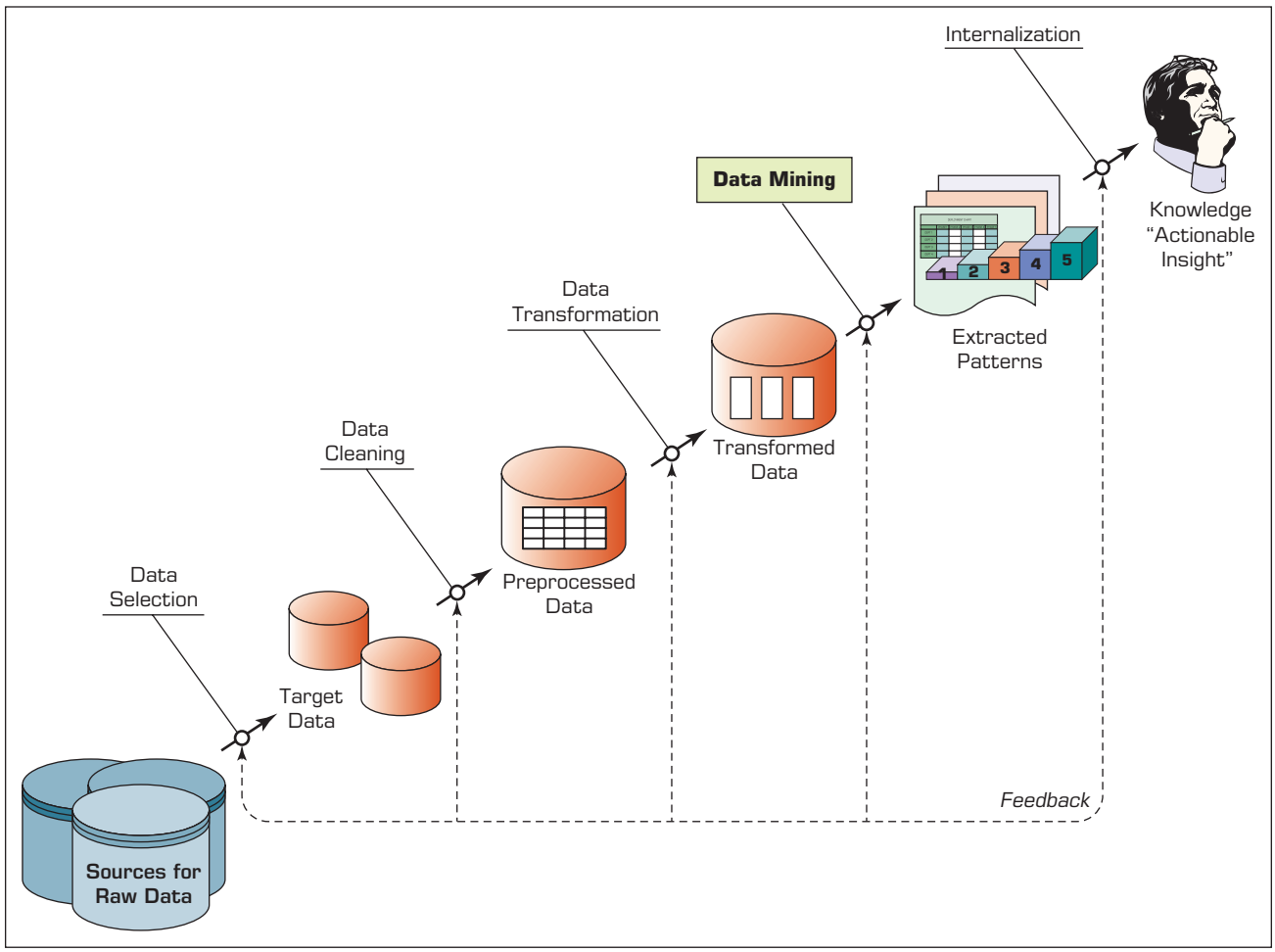


FIGURE 4.6 **KDD (Knowledge Discovery in Databases) Process.**

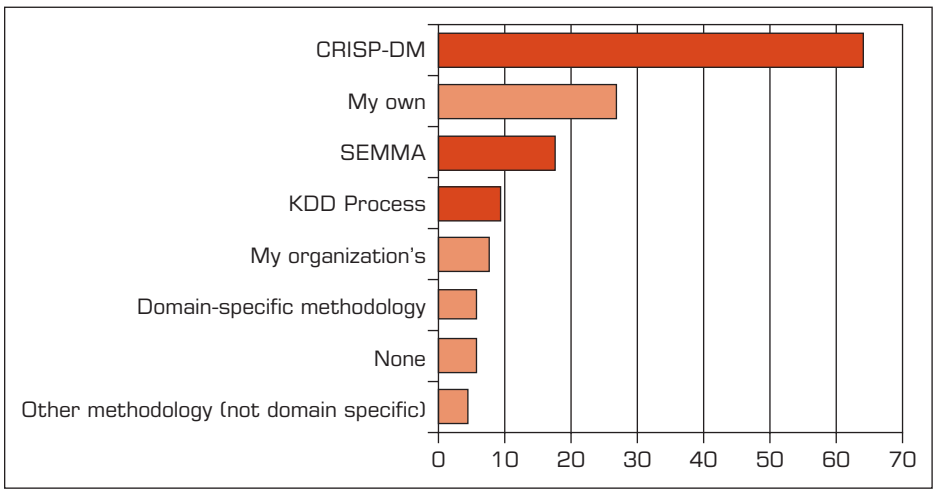


FIGURE 4.7 **Ranking of Data Mining Methodologies/Processes.** *Source:* Used with permission from **KDnuggets.com**.

SECTION 4.4 REVIEW QUESTIONS

1. What are the major data mining processes?
2. Why do you think the early phases (understanding of the business and understanding of the data) take the longest amount of time in data mining projects?
3. List and briefly define the phases in the CRISP-DM process.
4. What are the main data-preprocessing steps? Briefly describe each step and provide relevant examples.
5. How does CRISP-DM differ from SEMMA?

4.5 DATA MINING METHODS

Various methods are available for performing data mining studies, including classification, regression, clustering, and association. Most data mining software tools employ more than one technique (or algorithm) for each of these methods. This section describes the most popular data mining methods and explains their representative techniques.

Classification

Classification is perhaps the most frequently used data mining method for real-world problems. As a popular member of the machine-learning family of techniques, classification learns patterns from past data (a set of information—traits, variables, features—on characteristics of the previously labeled items, objects, or events) to place new instances (with unknown labels) into their respective groups or classes. For example, one could use classification to predict whether the weather on a particular day will be "sunny," "rainy," or "cloudy." Popular classification tasks include credit approval (i.e., good or bad credit risk), store location (e.g., good, moderate, bad), target marketing (e.g., likely customer, no hope), fraud detection (i.e., yes/no), and telecommunication (e.g., likely to turn to another phone company, yes/no). If what is being predicted is a class label (e.g., "sunny," "rainy," or "cloudy"), the prediction problem is called a *classification*; if it is a numeric value (e.g., temperature, such as 68°F), the prediction problem is called a **regression**.

Even though clustering (another popular data mining method) can also be used to determine groups (or class memberships) of things, there is a significant difference between the two. Classification learns the function between the characteristics of things (i.e., independent variables) and their membership (i.e., output variable) through a supervised learning process in which both types (input and output) of variables are presented to the algorithm; in clustering, the membership of the objects is learned through an unsupervised learning process by which only the input variables are presented to the algorithm. Unlike classification, clustering does not have a supervising (or controlling) mechanism that enforces the learning process; instead, clustering algorithms use one or more heuristics (e.g., multidimensional distance measure) to discover natural groupings of objects.

The most common two-step methodology of classification-type prediction involves model development/training and model testing/deployment. In the model development phase, a collection of input data, including the actual class labels, is used. After a model has been trained, the model is tested against the holdout sample for accuracy assessment and eventually deployed for actual use where it is to predict classes of new data instances (where the class label is unknown). Several factors are considered in assessing the model, including the following.

- **Predictive accuracy.** The model's ability to correctly predict the class label of new or previously unseen data. Prediction accuracy is the most commonly used assessment factor for classification models. To compute this measure, actual class

labels of a test data set are matched against the class labels predicted by the model. The accuracy can then be computed as the *accuracy rate,* which is the percentage of test data set samples correctly classified by the model (more on this topic is provided later in the chapter).

- **Speed.** The computational costs involved in generating and using the model where faster is deemed to be better.
- **Robustness.** The model's ability to make reasonably accurate predictions given noisy data or data with missing and erroneous values.
- **Scalability.** The ability to construct a prediction model efficiently given a rather large amount of data.
- **Interpretability.** The level of understanding and insight provided by the model (e.g., how and/or what the model concludes on certain predictions).

Estimating the True Accuracy of Classification Models

In classification problems, the primary source for accuracy estimation is the *confusion matrix* (also called a *classification matrix* or a *contingency table*). Figure 4.8 shows a confusion matrix for a two-class classification problem. The numbers along the diagonal from the upper left to the lower right represent correct decisions, and the numbers outside this diagonal represent the errors.

Table 4.1 provides equations for common accuracy metrics for classification models.

When the classification problem is not binary, the confusion matrix gets bigger (a square matrix with the size of the unique number of class labels), and accuracy metrics become limited to *per class accuracy rates* and the *overall classifier accuracy*.

$$(True\ \ Classification\ \ Rate)_i = \frac{(True\ \ Classification)}{\sum_{i=1}^{n}(False\ \ Classification)}$$

$$(Overall\ \ Classifier\ \ Accuracy)_i = \frac{\sum_{i=1}^{n}(Ture\ \ Classification)_i}{Total\ \ Number\ \ of\ \ Cases}$$

Estimating the accuracy of a classification model (or classifier) induced by a supervised learning algorithm is important for the following two reasons: First, it can be used

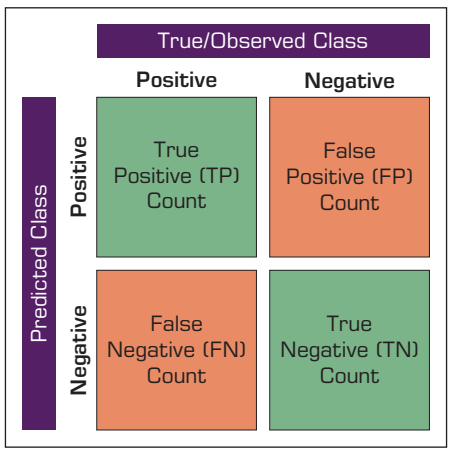


FIGURE 4.8 Simple Confusion Matrix for Tabulation of Two-Class Classification Results.

TABLE 4.1 Common Accuracy Metrics for Classification Models

Metric	Description
$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	The ratio of correctly classified instances (positives and negatives) divided by the total numbers of instances
$\text{True Positive Rate} = \frac{TP}{TP + FN}$	(a.k.a. Sensitivity) The ratio of correctly classified positives divided by the total positive count (i.e., hit rate or recall)
$\text{True Negative Rate} = \frac{TN}{TN + FP}$	(a.k.a. Specificity) The ratio of correctly classified negatives divided by the total negative count (i.e., false alarm rate)
$\text{Precision} = \frac{TP}{TP + FP}$	The ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified positives
$\text{Recall} = \frac{TP}{TP + FN}$	Ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified negatives

to estimate its future prediction accuracy, which could imply the level of confidence one should have in the classifier's output in the prediction system. Second, it can be used for choosing a classifier from a given set (identifying the "best" classification model among the many trained). The following are among the most popular estimation methodologies used for classification-type data mining models.

SIMPLE SPLIT The **simple split** (or holdout or test sample estimation) partitions the data into two mutually exclusive subsets called a *training set* and a *test set* (or *holdout set*). It is common to designate two-thirds of the data as the training set and the remaining one-third as the test set. The training set is used by the inducer (model builder), and the built classifier is then tested on the test set. An exception to this rule occurs when the classifier is an artificial neural network. In this case, the data are partitioned into three mutually exclusive subsets: training, validation, and testing. The validation set is used during model building to prevent overfitting. Figure 4.9 shows the simple split methodology.

The main criticism of this method is that it makes the assumption that the data in the two subsets are of the same kind (i.e., have the exact same properties). Because this is a simple random partitioning, in most realistic data sets where the data are skewed on the classification variable, such an assumption might not hold true. To improve this

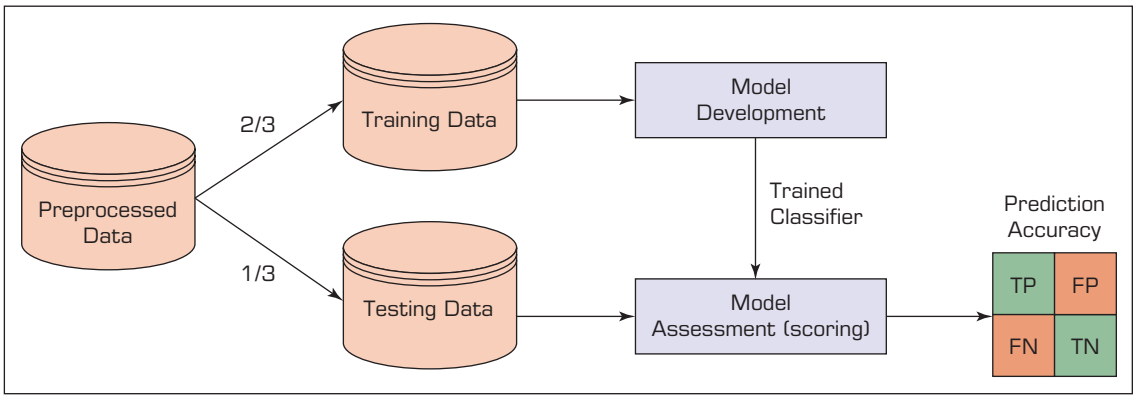


FIGURE 4.9 Simple Random Data Splitting.

situation, stratified sampling is suggested, where the strata become the output variable. Even though this is an improvement over the simple split, it still has a bias associated from the single random partitioning.

K-FOLD CROSS-VALIDATION To minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, one can use a methodology called ***k*-fold cross-validation**. In *k*-fold cross-validation, also called *rotation estimation,* the complete data set is randomly split into *k* mutually exclusive subsets of approximately equal size. The classification model is trained and tested *k* times. Each time it is trained on all but one fold and then tested on the remaining single fold. The cross-validation estimate of the overall accuracy of a model is calculated by simply averaging the *k* individual accuracy measures, as shown in the following equation:

$$\text{CVA} = \frac{1}{k}\sum_{i=1}^{k} A_i$$

where CVA stands for cross-validation accuracy, *k* is the number of folds used, and *A* is the accuracy measure (e.g., hit rate, sensitivity, specificity) of each fold. Figure 4.10 shows a graphical illustration of *k*-fold cross-validation where *k* is set to 10.

ADDITIONAL CLASSIFICATION ASSESSMENT METHODOLOGIES Other popular assessment methodologies include the following:

- ***Leave one out.*** The leave-one-out method is similar to the *k*-fold cross-validation where the *k* takes the value of 1; that is, every data point is used for testing once as many models are developed as there are data points. This is a time-consuming methodology, but sometimes for small data sets, it is a viable option.
- ***Bootstrapping.*** With **bootstrapping**, a fixed number of instances from the original data are sampled (with replacement) for training, and the rest of the data set is used for testing. This process is repeated as many times as desired.
- ***Jackknifing.*** Though similar to the leave-one-out methodology, with jackknifing, the accuracy is calculated by leaving one sample out at each iteration of the estimation process.
- ***Area under the ROC curve.*** The **area under the ROC curve** is a graphical assessment technique that plots the true positive rate on the *y*-axis and the false positive rate on the *x*-axis. The area under the ROC curve determines the accuracy

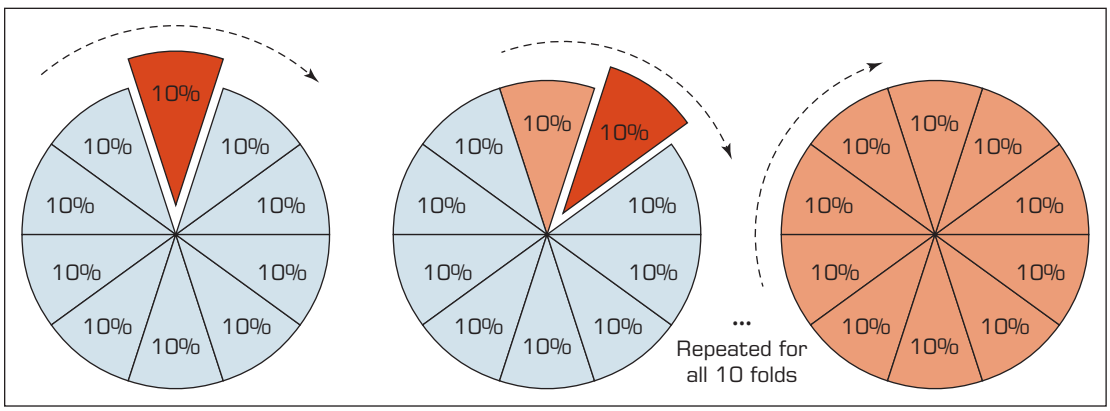


FIGURE 4.10 Graphical Depiction of *k*-Fold Cross-Validation.

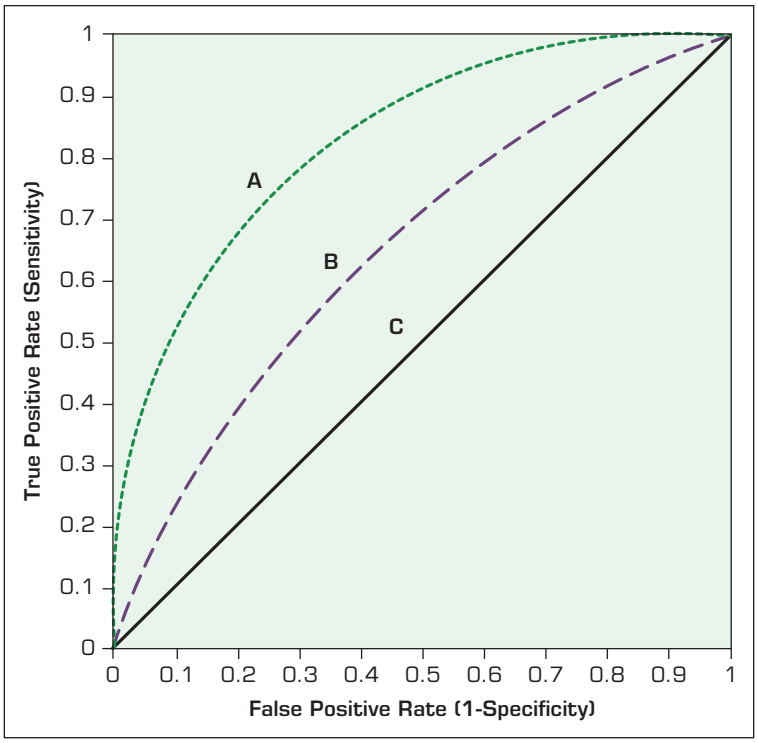


FIGURE 4.11 **Sample ROC Curve.**

measure of a classifier: A value of 1 indicates a perfect classifier; 0.5 indicates no better than random chance; in reality, the values would range between the two extreme cases. For example, in Figure 4.11, *A* has a better classification performance than *B*, whereas *C* is not any better than the random chance of flipping a coin.

Estimating the Relative Importance of Predictor Variables

Data mining methods (i.e., machine-learning algorithms) are really good at capturing complex relationships between input and output variables (producing very accurate prediction models) but are not nearly as good at explaining how they do what they do (i.e., model transparency). To mitigate this deficiency (also called the *black-box syndrome*), the machine-learning community proposed several methods, most of which are characterized as sensitivity analysis. In the context of predictive modeling, **sensitivity analysis** refers to an exclusive experimentation process aimed at discovering the cause-and-effect relationship between the input variables and output variable. Some of the variable importance methods are algorithm specific (i.e., applied to decision trees) and some are algorithm agnostic. Here are the most commonly used variable importance methods employed in machine learning and predictive modeling:

1. Developing and observing a well-trained decision tree model to see the relative discernibility of the input variables—the closer to the root of the tree a variable is used to split, the greater is its importance/relative-contribution to the prediction model.
2. Developing and observing a rich and large random forest model and assessing the variable split statistics. If the ratio of a given variable's selection into candidate counts (i.e., number of times a variable selected as the level-0 splitter is divided by number of times it was picked randomly as one of split candidates) is larger, its importance/relative-contribution is also greater.

3. Sensitivity analysis based on input value perturbation by which the input variables are gradually changed/perturbed one at a time and the relative change in the output is observed—the larger the change in the output, the greater the importance of the perturbed variable. This method is often used in feed-forward neural network modeling when all of the variables are numeric and standardized/normalized. Because this method is covered in Chapter 6 within the context of deep learning and deep neural networks, it is not be explained here.
4. Sensitivity analysis based on leave-one-out methodology. This method can be used for any type of predictive analytics method and therefore is further explained as follows.

The sensitivity analysis (based on leave-one-out methodology) relies on the experimental process of systematically removing input variables, one at a time, from the input variable set, developing and testing a model, and observing the impact of the absence of this variable on the predictive performance of the machine-learning model. The model is trained and tested (often using a *k*-fold cross validation) for each input variable (i.e., its absence in the input variable collection) to measure its contribution/importance to the model. A graphical depiction of the process is shown in Figure 4.12.

This method is often used for support vector machines, decision trees, logistic regression, and artificial neural networks. In his sensitivity analysis book, Saltelli (2002) formalized the algebraic representation of this measurement process:

$$S_i = \frac{V_i}{V(F_t)} = \frac{V(E(F_t|X_i))}{V(F_t)}$$

In the denominator of the equation, $V(F_t)$ refers to the variance in the output variable. In the numerator, $V(E(F_t|X_i))$, E is the expectation operator to call for an integral over parameter X_i; that is, inclusive of all input variables except X_i, the V, the variance operator, applies a further integral over X_i. The variable contribution (i.e., importance), represented as S_i for the i^{th} variable, is calculated as the normalized sensitivity measure. In a later study, Saltelli et al. (2004) proved that this equation is the most probable

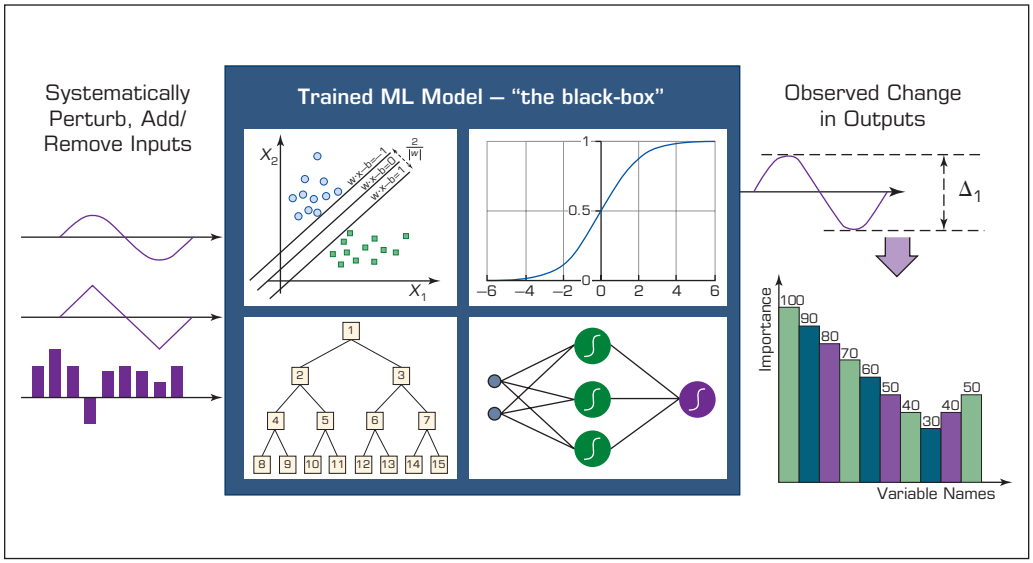


FIGURE 4.12 Graphical Depiction of the Sensitivity Analysis Process.

measure of model sensitivity that is capable of ranking input variables (i.e., the predictors) in the order of importance for any combination of interactions including the nonorthogonal relationships among the input variables. To properly combine the sensitivity analysis results for several prediction methods, one can use an information fusion–based methodology, particularly by modifying the preceding equation in such a way that the sensitivity measure of an input variable n obtained based on the information combined (i.e., fused) from m number of prediction models. The following equation represents this weighted summation function.

$$S_{n(fused)} = \sum_{i=1}^{m} \omega_i S_{in} = \omega_1 S_{1n} + \omega_2 S_{2n} + \ldots + \omega_m S_{mn}$$

In this equation, ω_i represents the normalized contribution/weight for each prediction model in which the level of contribution/weight of a model is calculated as a function of its relative predictive power—the larger the prediction power (i.e., accuracy) is, the higher is the value of ω.

CLASSIFICATION TECHNIQUES A number of techniques (or algorithms) are used for classification modeling, including the following:

- ***Decision tree analysis.*** Decision tree analysis (a machine-learning technique) is arguably the most popular classification technique in the data mining arena. A detailed description of this technique is given in the following section.
- ***Statistical analysis.*** Statistical techniques were the primary classification algorithm for many years until the emergence of machine-learning techniques. Statistical classification techniques include logistic regression and discriminant analysis, both of which make the assumptions that the relationships between the input and output variables are linear in nature, the data are normally distributed, and the variables are not correlated and are independent of each other. The questionable nature of these assumptions has led to the shift toward machine-learning techniques.
- ***Neural networks.*** These are among the most popular machine-learning techniques that can be used for classification-type problems.
- ***Case-based reasoning.*** This approach uses historical cases to recognize commonalities to assign a new case into the most probable category.
- ***Bayesian classifiers.*** This approach uses probability theory to build classification models based on the past occurrences that are capable of placing a new instance into a most probable class (or category).
- ***Genetic algorithms.*** This is the use of the analogy of natural evolution to build directed-search-based mechanisms to classify data samples.
- ***Rough sets.*** This method takes into account the partial membership of class labels to predefined categories in building models (collection of rules) for classification problems.

A complete description of all of these classification techniques is beyond the scope of this book; thus, only several of the most popular ones are presented here.

DECISION TREES Before describing the details of **decision trees**, we need to discuss some simple terminology. First, decision trees include many input variables that might have an impact on the classification of different patterns. These input variables are usually called *attributes*. For example, if we were to build a model to classify loan risks on the basis of just two characteristics—income and a credit rating—these two characteristics would be the attributes, and the resulting output would be the *class label* (e.g., low, medium, or high risk). Second, a tree consists of branches and nodes. A *branch* represents

the outcome of a test to classify a pattern using one of the attributes. A *leaf node* at the end represents the final class choice for a pattern (a chain of branches from the root node to the leaf node, which can be represented as a complex if-then statement).

The basic idea behind a decision tree is that it recursively divides a training set until each division consists entirely or primarily of examples from one class. Each nonleaf node of the tree contains a *split point*, which is a test of one or more attributes and determines how the data are to be divided further. Decision tree algorithms, in general, build an initial tree from training data such that each leaf node is pure, and they then prune the tree to increase its generalization, and hence, the prediction accuracy of test data.

In the growth phase, the tree is built by recursively dividing the data until each division is either pure (i.e., contains members of the same class) or relatively small. The basic idea is to ask questions whose answers would provide the most information, similar to what we do when playing the game "Twenty Questions."

The split used to partition the data depends on the type of the attribute used in the split. For a continuous attribute A, splits are of the form value(A) $< x$, where x is some "optimal" split value of A. For example, the split based on income could be "Income $<$ 50,000." For the categorical attribute A, splits are of the form that value(A) belongs to x where x is a subset of A. As an example, the split could be on the basis of gender: "Male versus Female."

A general algorithm for building a decision tree is as follows:

1. Create a root node and assign all of the training data to it.
2. Select the *best* splitting attribute.
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive (nonoverlapping) subsets along the lines of the specific split and move to the branches.
4. Repeat steps 2 and 3 for each and every leaf node until a stopping criterion is reached (e.g., the node is dominated by a single class label).

Many different algorithms have been proposed for creating decision trees. These algorithms differ primarily in terms of the way in which they determine the splitting attribute (and its split values), the order of splitting the attributes (splitting the same attribute only once or many times), the number of splits at each node (binary versus ternary), the stopping criteria, and the pruning of the tree (pre- versus postpruning). Some of the most well-known algorithms are ID3 (followed by C4.5 and C5 as the improved versions of ID3) from machine learning, classification and regression trees (CART) from statistics, and the chi-squared automatic interaction detector (CHAID) from pattern recognition.

When building a decision tree, the goal at each node is to determine the attribute and the split point of that attribute that best divides the training records to purify the class representation at that node. To evaluate the goodness of the split, some splitting indices have been proposed. Two of the most common ones are the Gini index and information gain. The Gini index is used in CART and Scalable PaRallelizable INduction of Decision Trees (SPRINT) algorithms. Versions of information gain are used in ID3 (and its newer versions, C4.5 and C5).

The **Gini index** has been used in economics to measure the diversity of a population. The same concept can be used to determine the purity of a specific class as the result of a decision to branch along a particular attribute or variable. The best split is the one that increases the purity of the sets resulting from a proposed split. Let us briefly look into a simple calculation of the Gini index.

If a data set S contains examples from n classes, the Gini index is defined as

$$gini(S) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is a relative frequency of class j in S. If a data set S is split into two subsets, S_1 and S_2 with sizes N_1 and N_2, respectively, the Gini index of the split data contains examples from n classes, and the Gini index is defined as

$$gini_{split}(S) = \frac{N_1}{N} gini(S_1) + \frac{N_2}{N} gini(S_2)$$

The attribute/split combination that provides the smallest $gini_{split}(S)$ is chosen to split the node. In such a determination, one should enumerate all possible splitting points for each attribute.

Information gain is the splitting mechanism used in ID3, which is perhaps the most widely known decision tree algorithm. It was developed by Ross Quinlan in 1986, and since then, he has evolved this algorithm into the C4.5 and C5 algorithms. The basic idea behind ID3 (and its variants) is to use a concept called *entropy* in place of the Gini index. **Entropy** measures the extent of uncertainty or randomness in a data set. If all the data in a subset belong to just one class, there is no uncertainty or randomness in that data set, so the entropy is zero. The objective of this approach is to build subtrees so that the entropy of each final subset is zero (or close to zero). Let us also look at the calculation of the information gain.

Assume that there are two classes: P (positive) and N (negative). Let the set of examples S contain p counts of class P and n counts of class N. The amount of information needed to decide if an arbitrary example in S belongs to P or N is defined as

$$I(p, n) = -\frac{p}{p+n} log_2 \frac{p}{p+n} - \frac{n}{p+n} log_2 \frac{n}{p+n}$$

Assume that using attribute A, the set S will be partitioned into sets $\{S_1, S_2, \ldots, g\}$. If S_i contains p_i examples of P and n_i examples of N, the entropy, or the expected information needed to classify objects in all subtrees, S_i, is

$$E(A) = \sum_{i=1}^{n} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Then the information that would be gained by branching on attribute A would be

$$Gain(A) = I(p,n) - E(A)$$

These calculations are repeated for each and every attribute, and the one with the highest information gain is selected as the splitting attribute. The basic ideas behind these splitting indices are rather similar, but the specific algorithmic details vary. A detailed definition of the ID3 algorithm and its splitting mechanism can be found in Quinlan (1986).

Application Case 4.5 illustrates how significant the gains can be if the right data mining techniques are used for a well-defined business problem.

Cluster Analysis for Data Mining

Cluster analysis is an essential data mining method for classifying items, events, or concepts into common groupings called *clusters*. The method is commonly used in biology, medicine, genetics, social network analysis, anthropology, archaeology, astronomy, character recognition, and even in management information systems (MIS) development. As data mining has increased in popularity, its underlying techniques have been applied to business, especially to marketing. Cluster analysis has been used extensively for fraud

Application Case 4.5

Influence Health Uses Advanced Predictive Analytics to Focus on the Factors That Really Influence People's Healthcare Decisions

Influence Health, Inc. provides the healthcare industry's only integrated digital consumer engagement and activation platform. It enables providers, employers, and payers to positively influence consumer decision making and health behaviors well beyond the physical care setting through personalized and interactive multichannel engagement. Since 1996, the Birmingham, Alabama–based company has helped more than 1,100 provider organizations influence consumers in a way that transforms financial and quality outcomes.

Healthcare is a personal business. Each patient's needs are different and require an individual response. On the other hand—as the cost of providing healthcare services continues to rise—hospitals and health systems increasingly need to harness economies of scale by catering to larger and larger populations. The challenge then becomes to provide a personalized approach while operating on a large scale. Influence Health specializes in helping its healthcare sector clients solve this challenge by getting to know their existing and potential patients better and targeting each individual with the appropriate health services at the right time. Advanced predictive analytics technology from IBM allows Influence Health to help its clients discover the factors that have the most influence on patients' healthcare decisions. By assessing the propensity of hundreds of millions of prospects to require specific healthcare services, Influence Health is able to boost revenues and response rates for healthcare campaigns, improving outcomes for its clients and their patients alike.

Targeting the Savvy Consumer

Today's healthcare industry is becoming more competitive than ever before. If the use of an organization's services drops, so do its profits. Rather than simply seeking out the nearest hospital or clinic, consumers are now more likely to make positive choices among healthcare providers. Paralleling efforts that are common in other industries, healthcare organizations must make increased efforts to market themselves effectively to both existing and potential patients, building long-term engagement and loyalty.

The keys to successful healthcare marketing are timeliness and relevance. If you can predict what kind of health services an individual prospect might need, you can engage and influence her or him much more effectively for wellness care.

Venky Ravirala, chief analytics officer at Influence Health, explains, "Healthcare organizations risk losing people's attention if they bombard them with irrelevant messaging. We help our clients avoid this risk by using analytics to segment their existing and potential prospects and market to them in a much more personal and relevant way."

Faster and More Flexible Analytics

As its client base has expanded, the total volume of data in Influence Health's analytics systems has grown to include over 195 million patient records with a detailed disease encounter history for several million patients. Ravirala comments, "With so much data to analyze, our existing method of scoring data was becoming too complex and time-consuming. We wanted to be able to extract insights at greater speed and accuracy."

By leveraging predictive analytics software from IBM, Influence Health is now able to develop models that calculate how likely each patient is to require particular services and express this likelihood as a percentage score. Microsegmentation and numerous disease-specific models draw on demographic, socioeconomic, geographical, behavioral, disease history, and census data and examine different aspects of each patient's predicted healthcare needs.

"The IBM solution allows us to combine all these models using an ensemble technique, which helps to overcome the limitations of individual models and provide more accurate results," comments Venky Ravirala, chief analytics officer at Influence Health. "It gives us the flexibility to apply multiple techniques to solve a problem and arrive at the best solution. It also automates much of the analytics process, enabling us to respond to clients' requests faster than before, and often give them a much deeper level of insight into their patient population."

For example, Influence Health decided to find out how disease prevalence and risk vary between different cohorts within the general population. By

(*Continued*)

Application Case 4.5 (Continued)

using very sophisticated cluster analysis techniques, the company was able to discover new comorbidity patterns that improve risk predictability for over 100 common diseases by up to 800 percent.

This helps to reliably differentiate between high-risk and very high-risk patients—making it easier to target campaigns at the patients and prospects who need them most. With insights like these in hand, Influence Health is able to use its healthcare marketing expertise to advise its clients on how best to allocate marketing resources.

"Our clients make significant budgeting decisions based on the guidance we give them," states Ravirala. "We help them maximize the impact of one-off campaigns—such as health insurance marketplace campaigns when Obamacare began—as well as their long-term strategic plans and ongoing marketing communications."

Reaching the Right Audience

By enabling its clients to target their marketing activities more effectively, Influence Health is helping to drive increased revenue and enhance population health. "Working with us, clients have been able to achieve return on investment of up to 12 to 1 through better targeted marketing," elaborates Ravirala. "And it's not just about revenues: by ensuring that vital healthcare information gets sent to the people who need it, we are helping our clients improve general health levels in the communities they serve."

Influence Health continues to refine its modeling techniques, gaining an ever-deeper understanding of the critical attributes that influence healthcare decisions. With a flexible analytics toolset at its fingertips, the company is well equipped to keep improving its service to clients. Ravirala explains, "In the future, we want to take our understanding of patient and prospect data to the next level, identifying patterns in behavior and incorporating analysis with machine-learning libraries. IBM SPSS has already given us the ability to apply and combine multiple models without writing a single line of code. We're eager to further leverage this IBM solution as we expand our healthcare analytics to support clinical outcomes and population health management services."

"We are achieving analytics on an unprecedented scale. Today, we can analyze 195 million records with 35 different models in less than two days—a task which was simply not possible for us in the past," says Ravirala.

Questions for Case 4.5

1. What does Influence Health do?
2. What were the company's challenges, proposed solutions, and obtained results?
3. How can data mining help companies in the healthcare industry (in ways other than the ones mentioned in this case)?

Source: Reprint Courtesy of International Business Machines Corporation, © (2018) International Business Machines Corporation.

detection (both credit card and e-commerce) and market segmentation of customers in contemporary CRM systems. More applications in business continue to be developed as the strength of cluster analysis is recognized and used.

Cluster analysis is an exploratory data analysis tool for solving classification problems. The objective is to sort cases (e.g., people, things, events) into groups, or clusters, so that the degree of association is strong among members of the same cluster and weak among members of different clusters. Each cluster describes the class to which its members belong. An obvious one-dimensional example of cluster analysis is to establish score ranges into which to assign class grades for a college class. This is similar to the cluster analysis problem that the U.S. Treasury faced when establishing new tax brackets in the 1980s. A fictional example of clustering occurs in J. K. Rowling's *Harry Potter* books. The Sorting Hat determines to which House (e.g., dormitory) to assign first-year students at the Hogwarts School. Another example involves determining how to seat guests at a wedding. As far as data mining goes, the importance of cluster analysis is that it can reveal

associations and structures in data that were not previously apparent but are sensible and useful once found.

Cluster analysis results can be used to

- Identify a classification scheme (e.g., types of customers).
- Suggest statistical models to describe populations.
- Indicate rules for assigning new cases to classes for identification, targeting, and diagnostic purposes.
- Provide measures of definition, size, and change in what were previously broad concepts.
- Find typical cases to label and represent classes.
- Decrease the size and complexity of the problem space for other data mining methods.
- Identify outliers in a specific domain (e.g., rare-event detection).

DETERMINING THE OPTIMAL NUMBER OF CLUSTERS Clustering algorithms usually require one to specify the number of clusters to find. If this number is not known from prior knowledge, it should be chosen in some way. Unfortunately, there is not an optimal way to calculate what this number is supposed to be. Therefore, several different heuristic methods have been proposed. The following are among the most commonly referenced ones:

- Look at the percentage of variance explained as a function of the number of clusters; that is, choose a number of clusters so that adding another cluster would not give much better modeling of the data. Specifically, if one graphs the percentage of variance explained by the clusters, there is a point at which the marginal gain will drop (giving an angle in the graph), indicating the number of clusters to be chosen.
- Set the number of clusters to $(n/2)^{1/2}$, where n is the number of data points.
- Use the Akaike information criterion (AIC), which is a measure of the goodness of fit (based on the concept of entropy), to determine the number of clusters.
- Use Bayesian information criterion, a model-selection criterion (based on maximum likelihood estimation), to determine the number of clusters.

ANALYSIS METHODS Cluster analysis might be based on one or more of the following general methods:

- Statistical methods (including both hierarchical and nonhierarchical), such as *k*-means or *k*-modes.
- Neural networks (with the architecture called *self-organizing map*).
- Fuzzy logic (e.g., fuzzy *c*-means algorithm).
- Genetic algorithms.

Each of these methods generally works with one of two general method classes:

- ***Divisive.*** With divisive classes, all items start in one cluster and are broken apart.
- ***Agglomerative.*** With agglomerative classes, all items start in individual clusters, and the clusters are joined together.

Most cluster analysis methods involve the use of a **distance measure** to calculate the closeness between pairs of items. Popular distance measures include Euclidian distance (the ordinary distance between two points that one would measure with a ruler) and Manhattan distance (also called the *rectilinear distance* or *taxicab distance*) between two points. Often, they are based on true distances that are measured, but this need not be so, as is typically the case in IS development. Weighted averages can be used to establish these distances. For example, in an IS development project, individual modules of

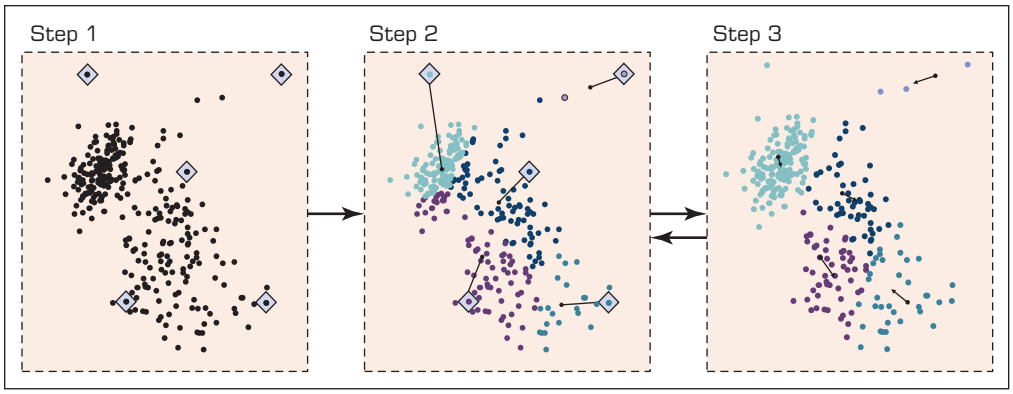


FIGURE 4.13 **Graphical Illustration of the Steps in the *k*-Means Algorithm.**

the system can be related by the similarity between their inputs, outputs, processes, and the specific data used. These factors are then aggregated, pairwise by item, into a single distance measure.

***K*-MEANS CLUSTERING ALGORITHM** The *k*-means algorithm (where *k* stands for the predetermined number of clusters) is arguably the most referenced clustering algorithm. It has its roots in traditional statistical analysis. As the name implies, the algorithm assigns each data point (customer, event, object, etc.) to the cluster whose center (also called the *centroid*) is the nearest. The center is calculated as the average of all the points in the cluster; that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The algorithm steps follow and are shown graphically in Figure 4.13:

Initialization step: Choose the number of clusters (i.e., the value of *K*).

Step 1: Randomly generate *k* random points as initial cluster centers.
Step 2: Assign each point to the nearest cluster center.
Step 3: Recompute the new cluster centers.

Repetition step: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

Association Rule Mining

Association rule mining (also known as *affinity analysis* or *market-basket analysis*) is a popular data mining method that is commonly used as an example to explain what data mining is and what it can do to a technologically less-savvy audience. Most of you might have heard the famous (or infamous, depending on how you look at it) relationship discovered between the sales of beer and diapers at grocery stores. As the story goes, a large supermarket chain (maybe Walmart, maybe not; there is no consensus on which supermarket chain it was) did an analysis of customers' buying habits and found a statistically significant correlation between purchases of beer and purchases of diapers. It was theorized that the reason for this was that fathers (presumably young men) were stopping off at the supermarket to buy diapers for their babies (especially on Thursdays), and because they could no longer go down to the sports bar as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have placed the diapers next to the beer, resulting in increased sales of both.

In essence, association rule mining aims to find interesting relationships (affinities) between variables (items) in large databases. Because of its successful application to retail business problems, it is commonly called *market-basket analysis*. The main idea

in market-basket analysis is to identify strong relationships among different products (or services) that are usually purchased together (show up in the same basket together, either a physical basket at a grocery store or a virtual basket at an e-commerce Web site). For example, 65 percent of those who buy comprehensive automobile insurance also buy health insurance; 80 percent of those who buy books online also buy music online; 60 percent of those who have high blood pressure and are overweight have high cholesterol; 70 percent of the customers who buy a laptop computer and virus protection software also buy extended service plans.

The input to market-basket analysis is the simple point-of-sale transaction data when a number of products and/or services purchased together (just like the content of a purchase receipt) are tabulated under a single transaction instance. The outcome of the analysis is invaluable information that can be used to better understand customer-purchase behavior to maximize the profit from business transactions. A business can take advantage of such knowledge by (1) putting the items next to each other to make it more convenient for the customers to pick them up together and not forget to buy one when buying the others (increasing sales volume), (2) promoting the items as a package—do not put one on sale if the other(s) are on sale, and (3) placing them apart from each other so that the customer has to walk the aisles to search for it, and by doing so, potentially seeing and buying other items.

Applications of market-basket analysis include cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration. In essence, market-basket analysis helps businesses infer customer needs and preferences from their purchase patterns. Outside the business realm, association rules are successfully used to discover relationships between symptoms and illnesses, diagnosis and patient characteristics and treatments (which can be used in a medical decision support system), and genes and their functions (which can be used in genomics projects), among others. Here are a few common areas and uses for association rule mining:

- ***Sales transactions:*** Combinations of retail products purchased together can be used to improve product placement on the sales floor (placing products that go together in close proximity) and promotional pricing of products (not having promotions on both products that are often purchased together).
- ***Credit card transactions:*** Items purchased with a credit card provide insight into other products the customer is likely to purchase or fraudulent use of credit card numbers.
- ***Banking services:*** The sequential patterns of services used by customers (checking account followed by savings account) can be used to identify other services they might be interested in (investment account).
- ***Insurance service products:*** Bundles of insurance products bought by customers (car insurance followed by home insurance) can be used to propose additional insurance products (life insurance), or unusual combinations of insurance claims can be a sign of fraud.
- ***Telecommunication services:*** Commonly purchased groups of options (e.g., call waiting, caller ID, three-way calling) help better structure product bundles to maximize revenue; the same is also applicable to multichannel telecom providers with phone, television, and Internet service offerings.
- ***Medical records:*** Certain combinations of conditions can indicate increased risk of various complications; or, certain treatment procedures at certain medical facilities can be tied to certain types of infections.

A good question to ask with respect to the patterns/relationships that association rule mining can discover is "Are all association rules interesting and useful?" To answer

such a question, association rule mining uses two common metrics: **support**, and **confidence** and **lift**. Before defining these terms, let's get a little technical by showing what an association rule looks like:

$$X \Rightarrow Y \ [Supp(\%),\ Conf(\%)]$$
$$\{\text{Laptop Computer, Antivirus software}\} \Rightarrow \{\text{Extended Service Plan}\}\ [30\%, 70\%]$$

Here, X (products and/or service—called the *left-hand side, LHS,* or *antecedent*) is associated with Y (products and/or service—called the *right-hand side, RHS,* or *consequent*). S is the support, and C is the confidence for this particular rule. Here are the simple formulas for *Supp, Conf,* and *Lift*.

$$Support = Supp(X \Rightarrow Y) = \frac{Number\ of\ Baskets\ that\ contains\ both\ X\ and\ Y}{Total\ Number\ of\ Baskets}$$

$$Confidence = Conf(X \Rightarrow Y) = \frac{Supp(X \Rightarrow Y)}{Supp(X)}$$

$$Lift(X \Rightarrow Y) = \frac{Conf(X \Rightarrow Y)}{Expected\ Conf(X \Rightarrow Y)} = \frac{\frac{S(X \Rightarrow Y)}{S(X)}}{\frac{S(X) * S(Y)}{S(X)}} = \frac{S(X \Rightarrow Y)}{S(X) * S(Y)}$$

The support (S) of a collection of products is the measure of how often these products and/or services (i.e., LHS + RHS = Laptop Computer, Antivirus Software, and Extended Service Plan) appear together in the same transaction; that is, the proportion of transactions in the data set that contain all of the products and/or services mentioned in a specific rule. In this example, 30 percent of all transactions in the hypothetical store database had all three products present in a single sales ticket. The confidence of a rule is the measure of how often the products and/or services on the RHS (consequent) go together with the products and/or services on the LHS (antecedent), that is, the proportion of transactions that include LHS while also including the RHS. In other words, it is the conditional probability of finding the RHS of the rule present in transactions where the LHS of the rule already exists. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence of a rule is defined as the product of the support values of the LHS and the RHS divided by the support of the LHS.

Several algorithms are available for discovering association rules. Some well-known algorithms include Apriori, Eclat, and FP-Growth. These algorithms only do half the job, which is to identify the frequent itemsets in the database. Once the frequent itemsets are identified, they need to be converted into rules with antecedent and consequent parts. Determination of the rules from frequent itemsets is a straightforward matching process, but the process can be time consuming with large transaction databases. Even though there can be many items on each section of the rule, in practice the consequent part usually contains a single item. In the following section, one of the most popular algorithms for identification of frequent itemsets is explained.

APRIORI ALGORITHM The **Apriori algorithm** is the most commonly used algorithm to discover association rules. Given a set of itemsets (e.g., sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets that are common to at least a minimum number of the itemsets (i.e., complies with a minimum support). Apriori uses a bottom-up approach by which frequent subsets are extended

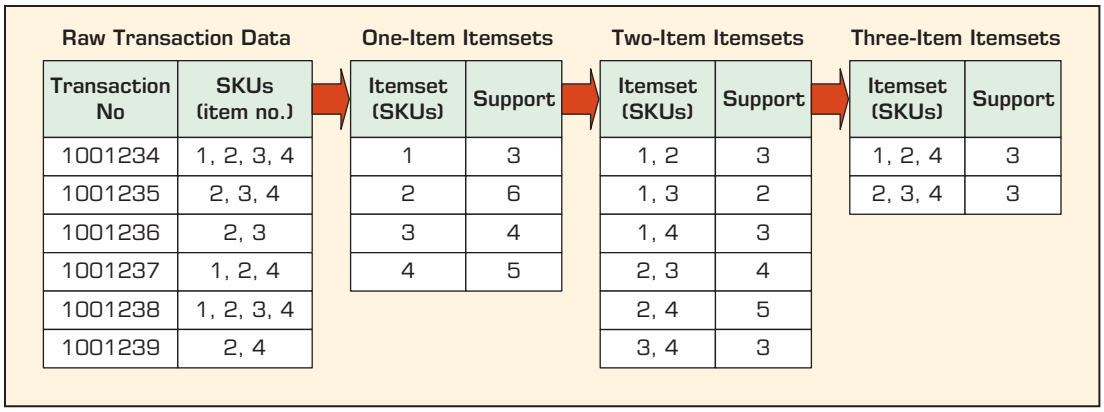


FIGURE 4.14 **Identification of Frequent Itemsets in the Apriori Algorithm.**

one item at a time (a method known as *candidate generation* by which the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, etc.), and groups of candidates at each level are tested against the data for minimum support. The algorithm terminates when no further successful extensions are found.

As an illustrative example, consider the following. A grocery store tracks sales transactions by SKU (stock-keeping unit) and thus knows which items are typically purchased together. The database of transactions along with the subsequent steps in identifying the frequent itemsets is shown in Figure 4.14. Each SKU in the transaction database corresponds to a product, such as "1 = butter," "2 = bread," "3 = water," and so on. The first step in Apriori is to count the frequencies (i.e., the supports) of each item (one-item itemsets). For this overly simplified example, let us set the minimum support to 3 (or 50, meaning an itemset is considered to be a frequent itemset if it shows up in at least 3 of 6 transactions in the database). Because all the one-item itemsets have at least 3 in the support column, they are all considered frequent itemsets. However, had any of the one-item itemsets not been frequent, they would not have been included as a possible member of possible two-item pairs. In this way, Apriori *prunes* the tree of all possible itemsets. As Figure 4.14 shows, using one-item itemsets, all possible two-item itemsets are generated and the transaction database is used to calculate their support values. Because the two-item itemset {1, 3} has a support less than 3, it should not be included in the frequent itemsets that will be used to generate the next-level itemsets (three-item itemsets). The algorithm seems deceivingly simple, but only for small data sets. In much larger data sets, especially those with huge amounts of items present in low quantities and small amounts of items present in big quantities, the search and calculation become a computationally intensive process.

SECTION 4.5 REVIEW QUESTIONS

1. Identify at least three of the main data mining methods.
2. Give examples of situations in which classification would be an appropriate data mining technique. Give examples of situations in which regression would be an appropriate data mining technique.
3. List and briefly define at least two classification techniques.
4. What are some of the criteria for comparing and selecting the best classification technique?

5. Briefly describe the general algorithm used in decision trees.
6. Define *Gini index*. What does it measure?
7. Give examples of situations in which cluster analysis would be an appropriate data mining technique.
8. What is the major difference between cluster analysis and classification?
9. What are some of the methods for cluster analysis?
10. Give examples of situations in which association would be an appropriate data mining technique.

4.6 DATA MINING SOFTWARE TOOLS

Many software vendors provide powerful data mining tools. Examples of these vendors include IBM (IBM SPSS Modeler, formerly known as SPSS PASW Modeler and Clementine), SAS (Enterprise Miner), Dell (Statistica, formerly known as StatSoft Statistica Data Miner), SAP (Infinite Insight, formerly known as KXEN Infinite Insight), Salford Systems (CART, MARS, TreeNet, RandomForest), Angoss (KnowledgeSTUDIO, KnowledgeSEEKER), and Megaputer (PolyAnalyst). Noticeably but not surprisingly, the most popular data mining tools are developed by the well-established statistical software companies (SAS, SPSS, and StatSoft)—largely because statistics is the foundation of data mining, and these companies have the means to cost-effectively develop them into full-scale data mining systems. Most of the business intelligence tool vendors (e.g., IBM Cognos, Oracle Hyperion, SAP Business Objects, Tableau, Tibco, Qlik, MicroStrategy, Teradata, and Microsoft) also have some level of data mining capabilities integrated into their software offerings. These BI tools are still primarily focused on multidimensional modeling and data visualization and are not considered to be direct competitors of the data mining tool vendors.

In addition to these commercial tools, several open source and/or free data mining software tools are available online. Traditionally, especially in educational circles, the most popular free and open source data mining tool is **Weka**, which was developed by a number of researchers from the University of Waikato in New Zealand (the tool can be downloaded from **cs.waikato.ac.nz/ml/weka**). Weka includes a large number of algorithms for different data mining tasks and has an intuitive user interface. Recently, a number of free open source, highly capable data mining tools emerged: leading the pack are **KNIME** (**knime.org**) and **RapidMiner** (**rapidminer.com**). Their graphically enhanced user interfaces, employment of a rather large number of algorithms, and incorporation of a variety of data visualization features set them apart from the rest of the free tools. These two free software tools are also platform agnostic (i.e., can natively run on both Windows and Mac operating systems). With a recent change in its offerings, RapidMiner has created a scaled-down version of its analytics tool for free (i.e., community edition) while making the full commercial product. Therefore, once listed under the free/open source tools category, RapidMiner today is often listed under commercial tools. The main difference between commercial tools, such as SAS Enterprise Miner, IBM SPSS Modeler, and Statistica, and free tools, such as Weka, RapidMiner (community edition), and KNIME, is the computational efficiency. The same data mining task involving a rather large and feature-rich data set can take much longer to complete with the free software tools, and for some algorithms, the job might not even be completed (i.e., crashing due to the inefficient use of computer memory). Table 4.2 lists a few of the major products and their Web sites.

A suite of business intelligence and analytics capabilities that has become increasingly more popular for data mining studies is **Microsoft's SQL Server** (it has included

TABLE 4.2 Selected Data Mining Software

Product Name	Web Site (URL)
IBM SPSS Modeler	**www-01.ibm.com/software/analytics/spss/products/modeler/**
IBM Watson Analytics	**ibm.com/analytics/watson-analytics/**
SAS Enterprise Miner	**sas.com/en_id/software/analytics/enterprise-miner.html**
Dell Statistica	**statsoft.com/products/statistica/product-index**
PolyAnalyst	**megaputer.com/site/polyanalyst.php**
CART, RandomForest	**salford-systems.com**
Insightful Miner	**solutionmetrics.com.au/products/iminer/default.html**
XLMiner	**solver.com/xlminer-data-mining**
SAP InfiniteInsight (KXEN)	**help.sap.com/ii**
GhostMiner	**qs.pl/ghostminer**
SQL Server Data Mining	**msdn.microsoft.com/en-us/library/bb510516.aspx**
Knowledge Miner	**knowledgeminer.com**
Teradata Warehouse Miner	**teradata.com/products-and-services/teradata-warehouse-miner/**
Oracle Data Mining (ODM)	**oracle.com/technetwork/database/options/odm/**
FICO Decision Management	**fico.com/en/analytics/decision-management-suite/**
Orange Data Mining Tool	**orange.biolab.si/**
Zementis Predictive Analytics	**zementis.com**

increasingly more analytics capabilities, such as BI and predictive modeling modules, starting with the SQL Server 2012 version) where data and the models are stored in the same relational database environment, making model management a considerably easier task. **Microsoft Enterprise Consortium** serves as the worldwide source for access to Microsoft's SQL Server software suite for academic purposes—teaching and research. The consortium has been established to enable universities around the world to access enterprise technology without having to maintain the necessary hardware and software on their own campus. The consortium provides a wide range of business intelligence development tools (e.g., data mining, cube building, business reporting) as well as a number of large, realistic data sets from Sam's Club, Dillard's, and Tyson Foods. The Microsoft Enterprise Consortium is free of charge and can be used only for academic purposes. The Sam M. Walton College of Business at the University of Arkansas hosts the enterprise system and allows consortium members and their students to access these resources using a simple remote desktop connection. The details about becoming a part of the consortium along with easy-to-follow tutorials and examples can be found at **walton.uark.edu/enterprise/**.

In May 2016, **KDnuggets.com** conducted the 13th Annual Software Poll on the following question: "What software have you used for Analytics, Data Mining, Data Science, Machine Learning projects in the past 12 months?" The poll received remarkable participation from analytics and data science community and vendors, attracting 2,895

voters, who chose from a record number of 102 different tools. Here are some of the interesting findings that came from the poll:

- R remains the leading tool, with 49 percent shares (up from 46.9% in 2015), but Python usage grew faster and almost caught up to R with 45.8 percent shares (up from 30.3%).
- RapidMiner remains the most popular general platform for data mining/data science with about 33 percent shares. Notable tools with the most growth in popularity include g, Dataiku, MLlib, H2O, Amazon Machine Learning, scikit-learn, and IBM Watson.
- The increased choice of tools is reflected in wider usage. The average number of tools used was 6.0 (versus 4.8 in May 2015).
- The usage of Hadoop/Big Data tools increased to 39 percent up from 29 percent in 2015 (and 17% in 2014) driven by Apache Spark, MLlib (Spark Machine Learning Library), and H2O.
- The participation by region was United States/Canada (40%), Europe (39%), Asia (9.4%), Latin America (5.8%), Africa/MidEast (2.9%), and Australia/New Zealand (2.2%).
- This year, 86 percent of voters used commercial software, and 75 percent used free software. About 25 percent used only commercial software, and 13 percent used only open source/free software. A majority of 61 percent used both free and commercial software, similar to 64 percent in 2015.
- The use of Hadoop/Big Data tools increased to 39 percent, up from 29 percent in 2015 and 17 percent in 2014, driven mainly by big growth in Apache Spark, MLlib (Spark Machine Learning Library), and H2O, which we include among Big Data tools.
- For the second year, **KDnuggets.com**'s poll included Deep Learning tools. This year, 18 percent of voters used Deep Learning tools, doubling the 9 percent in 2015—Google Tensorflow jumped to first place, displacing last year's leader, Theano/Pylearn2 ecosystem.
- In the programming languages category, Python, Java, Unix tools, and Scala grew in popularity, while C/C++, Perl, Julia, F#, Clojure, and Lisp declined.

To reduce bias through multiple voting, in this poll **KDnuggets.com** used e-mail verification and, by doing so, aimed to make results more representative of the reality in the analytics world. The results for the top 40 software tools (as per total number of votes received) are shown in Figure 4.15. The horizontal bar chart also makes a differentiation among free/open source, commercial, and Big Data/Hadoop tools using a color-coding schema.

Application Case 4.6 is about a research study in which a number of software tools and data mining techniques were used to build data mining models to predict financial success (box-office receipts) of Hollywood movies while they are nothing more than ideas.

SECTION 4.6 REVIEW QUESTIONS

1. What are the most popular commercial data mining tools?
2. Why do you think the most popular tools are developed by statistics-based companies?
3. What are the most popular free data mining tools? Why are they gaining overwhelming popularity (especially R)?
4. What are the main differences between commercial and free data mining software tools?
5. What would be your top five selection criteria for a data mining tool? Explain.

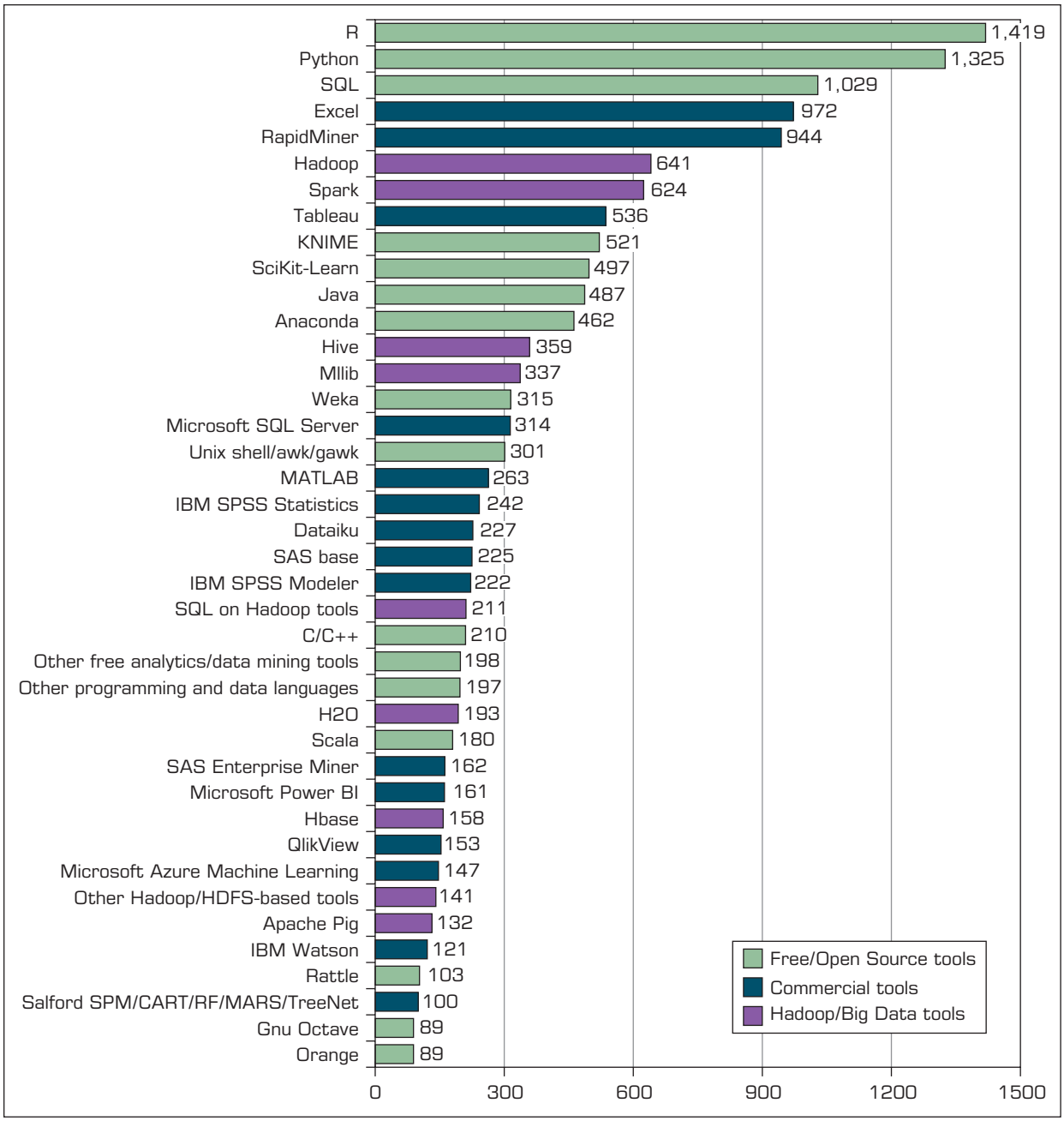


FIGURE 4.15 **Popular Data Mining Software Tools (Poll Results).** *Source:* Used with permission from **KDnuggets.com**.

Application Case 4.6

Data Mining goes to Hollywood: Predicting Financial Success of Movies

Predicting box-office receipts (i.e., financial success) of a particular motion picture is an interesting and challenging problem. According to some domain experts, the movie industry is the "land of hunches and wild guesses" due to the difficulty associated with forecasting product demand, making the movie

(*Continued*)

Application Case 4.6 (Continued)

business in Hollywood a risky endeavor. In support of such observations, Jack Valenti (the longtime president and CEO of the Motion Picture Association of America) once mentioned that "no one can tell you how a movie is going to do in the marketplace . . . not until the film opens in darkened theatre and sparks fly up between the screen and the audience." Entertainment industry trade journals and magazines have been full of examples, statements, and experiences that support such a claim.

Like many other researchers who have attempted to shed light on this challenging real-world problem, Ramesh Sharda and Dursun Delen have been exploring the use of data mining to predict the financial performance of a motion picture at the box office before it even enters production (while the movie is nothing more than a conceptual idea). In their highly publicized prediction models, they convert the forecasting (or regression) problem into a classification problem; that is, rather than forecasting the point estimate of box-office receipts, they classify a movie based on its box-office receipts in one of nine categories, ranging from "flop" to "blockbuster," making the problem a multinomial classification problem. Table 4.3 illustrates the definition of the nine classes in terms of the range of box-office receipts.

Data

Data were collected from a variety of movie-related databases (e.g., ShowBiz, IMDb, IMSDb, AllMovie, BoxofficeMojo) and consolidated into a single data set. The data set for the most recently developed models contained 2,632 movies released between 1998 and 2006. A summary of the independent variables along with their specifications is provided in Table 4.4. For more descriptive details and justification for inclusion of these independent variables, the reader is referred to Sharda and Delen (2006).

The Methodology

Using a variety of data mining methods, including neural networks, decision trees, SVMs, and three types of ensembles, Sharda and Delen (2006) developed the prediction models. The data from 1998 to 2005 were used as training data to build the prediction models, and the data from 2006 were used as the test data to assess and compare the models' prediction accuracy. Figure 4.16 shows a screenshot of IBM SPSS Modeler (formerly Clementine data mining tool) depicting the process map employed for the prediction problem. The upper-left side of the process map shows the model development

TABLE 4.3 Movie Classification based on Receipts

Class No.	1	2	3	4	5	6	7	8	9
Range (in millions of dollars)	>1 (Flop)	>1 <6 10	>10 <20	>20 <6 40	>40 <6 65	>65 <6 100	>100 <6 150	>150 <6 200	>200 (Blockbuster)

TABLE 4.4 Summary of Independent Variables

Independent Variable	Number of Values	Possible Values
MPAA Rating	5	G, PG, PG-13, R, NR
Competition	3	High, medium, low
Star value	3	High, medium, low
Genre	10	Sci-Fi, Historic Epic Drama, Modern Drama, Politically Related, Thriller, Horror, Comedy, Cartoon, Action, Documentary
Special effects	3	High, medium, low
Sequel	2	Yes, no
Number of screens	1	A positive integer between 1 and 3,876

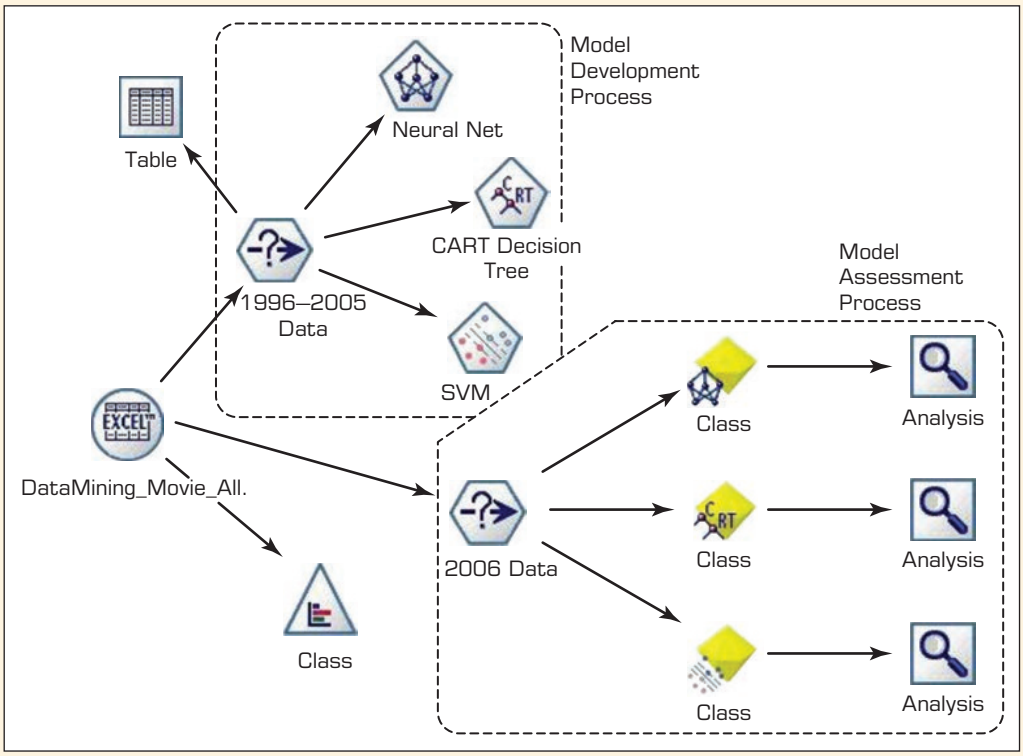


FIGURE 4.16 **Process Flow Screenshot for the Box-Office Prediction System.** *Source:* Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

process, and the lower-right corner of the process map shows the model assessment (i.e., testing or scoring) process (more details on the IBM SPSS Modeler tool and its usage can be found on the book's Web site).

The Results

Table 4.5 provides the prediction results of all three data mining methods as well as the results of the three different ensembles. The first performance measure is the percentage of correct classification rate, which is called *Bingo*. Also reported in the table is the *1-Away* correct classification rate (i.e., within one category). The results indicate that SVM performed the best among the individual prediction models followed by ANN; the worst of the three was the CART decision tree algorithm. In general, the ensemble models performed better than the individual prediction models of which the fusion algorithm performed the best. What is probably more important to decision makers and standing out in the results table is the significantly low standard deviation obtained from the ensembles compared to the individual models.

The Conclusion

The researchers claim that these prediction results are better than any reported in the published literature for this problem domain. Beyond the attractive accuracy of their prediction results of the box-office receipts, these models could also be used to further analyze (and potentially optimize) the decision variables to maximize the financial return. Specifically, the parameters used for modeling could be altered using the already trained prediction models to better understand the impact of different parameters on the end results. During this process, which is commonly referred to as *sensitivity analysis*, the decision maker of a given entertainment firm could find out, with a fairly high accuracy level, how much value a specific actor (or a specific release date, or the addition of more technical effects, etc.) brings to the financial success of a film, making the underlying system an invaluable decision aid.

(*Continued*)

Application Case 4.6 (Continued)

TABLE 4.5 Tabulated Prediction Results for Individual and Ensemble Models

	Prediction Models					
	Individual Models			Ensemble Models		
Performance Measure	SVM	ANN	CART	Random Forest	Boosted Tree	Fusion (average)
Count (Bingo)	192	182	140	189	187	**194**
Count (1-Away)	104	120	126	121	104	**120**
Accuracy (% Bingo)	55.49%	52.60%	40.46%	54.62%	54.05%	**56.07%**
Accuracy (% 1-Away)	85.55%	87.28%	76.88%	89.60%	84.10%	**90.75%**
Standard deviation	0.93	0.87	1.05	0.76	0.84	**0.63**

Questions for Case 4.6

1. Why is it important for many Hollywood professionals to predict the financial success of movies?
2. How can data mining be used for predicting financial success of movies before the start of their production process?
3. How do you think Hollywood performed, and perhaps still is performing, this task without the help of data mining tools and techniques?

Sources: R. Sharda & D. Delen, "Predicting Box-Office Success of Motion Pictures with Neural Networks," *Expert Systems with Applications, 30*, 2006, pp. 243–254; D. Delen, R. Sharda, & P. Kumar, "Movie Forecast Guru: A Web-Based DSS for Hollywood Managers," *Decision Support Systems, 43*(4), 2007, pp. 1151–1170.

4.7 DATA MINING PRIVACY ISSUES, MYTHS, AND BLUNDERS

Data that are collected, stored, and analyzed in data mining often contain information about real people. Such information can include identification data (name, address, Social Security number, driver's license number, employee number, etc.), demographic data (e.g., age, sex, ethnicity, marital status, number of children), financial data (e.g., salary, gross family income, checking or savings account balance, home ownership, mortgage or loan account specifics, credit card limits and balances, investment account specifics), purchase history (i.e., what is bought from where and when—either from vendor's transaction records or from credit card transaction specifics), and other personal data (e.g., anniversary, pregnancy, illness, loss in the family, bankruptcy filings). Most of these data can be accessed through some third-party data providers. The main question here is the privacy of the person to whom the data belong. To maintain the privacy and protection of individuals' rights, data mining professionals have ethical (and often legal) obligations. One way to accomplish this is the process of de-identification of the customer records prior to applying data mining applications so that the records cannot be traced to an individual. Many publicly available data sources (e.g., CDC data, SEER data, UNOS data) are already de-identified. Prior to accessing these data sources, users are often asked to consent that under no circumstances will they try to identify the individuals behind those figures.

There have been a number of instances in the recent past when companies shared their customer data with others without seeking the explicit consent of their customers. For instance, as most of you might recall, in 2003, JetBlue Airlines provided more than 1 million passenger records of customers to Torch Concepts, a U.S. government contractor. Torch

then subsequently augmented the passenger data with additional information such as family sizes and Social Security numbers—information purchased from the data broker Acxiom. The consolidated personal database was intended to be used for a data mining project to develop potential terrorist profiles. All of this was done without notification or consent of passengers. When news of the activities got out, however, dozens of privacy lawsuits were filed against JetBlue, Torch, and Acxiom, and several U.S. senators called for an investigation into the incident (Wald, 2004). Similar, but not as dramatic, privacy-related news was reported in the recent past about popular social network companies that allegedly were selling customer-specific data to other companies for personalized target marketing.

Another peculiar story about privacy concerns made it to the headlines in 2012. In this instance, the company, Target, did not even use any private and/or personal data. Legally speaking, there was no violation of any laws. The story is summarized in Application Case 4.7.

Application Case 4.7 Predicting Customer Buying Patterns—The Target Story

In early 2012, an infamous story appeared concerning Target's practice of predictive analytics. The story was about a teenage girl who was being sent advertising flyers and coupons by Target for the kinds of things that a mother-to-be would buy from a store like Target. The story goes like this: An angry man went into a Target outside of Minneapolis, demanding to talk to a manager: "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?" The manager had no idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture, and pictures of smiling infants. The manager apologized and then called a few days later to apologize again. On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

As it turns out, Target figured out a teen girl was pregnant before her father did! Here is how the company did it. Target assigns every customer a Guest ID number (tied to his or her credit card, name, or e-mail address) that becomes a placeholder that keeps a history of everything the person has bought. Target augments these data with any demographic information that it had collected from the customer or had bought from other information sources. Using this information, Target looked at historical buying data for all the females who had signed up for Target baby registries in the past. They analyzed the data from all directions, and soon enough, some useful patterns emerged. For example, lotions and special vitamins were among the products with interesting purchase patterns. Lots of people buy lotion, but what an analyst noticed was that women on the baby registry were buying larger quantities of unscented lotion around the beginning of their second trimester. Another analyst noted that sometime in the first 20 weeks, pregnant women loaded up on supplements like calcium, magnesium, and zinc. Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-large bags of cotton balls, in addition to hand sanitizers and washcloths, it signals that they could be getting close to their delivery date. In the end, the analysts were able to identify about 25 products that, when analyzed together, allowed them to assign each shopper a "pregnancy prediction" score. More important, they could also estimate a woman's due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.

If you look at this practice from a legal perspective, you would conclude that Target did not use any information that violates customer privacy; rather, they used transactional data that almost every other retail chain is collecting and storing (and perhaps analyzing) about their customers. What was disturbing in this scenario was perhaps the targeted concept: pregnancy. Certain events or concepts should

(*Continued*)

Application Case 4.7 (Continued)

be off limits or treated extremely cautiously, such as terminal disease, divorce, and bankruptcy.

Questions for Case 4.7

1. What do you think about data mining and its implication for privacy? What is the threshold between discovery of knowledge and infringement of privacy?
2. Did Target go too far? Did it do anything illegal? What do you think Target should have done? What do you think Target should do next (quit these types of practices)?

Sources: K. Hill, "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did," *Forbes,* February 16, 2012; R. Nolan, "Behind the Cover Story: How Much Does Target Know?", February 21, 2012. **NYTimes.com**.

Data Mining Myths and Blunders

Data mining is a powerful analytical tool that enables business executives to advance from describing the nature of the past (looking at a rearview mirror) to predicting the future (looking ahead) to better manage their business operations (making accurate and timely decisions). Data mining helps marketers find patterns that unlock the mysteries of customer behavior. The results of data mining can be used to increase revenue and reduce cost by identifying fraud and discovering business opportunities, offering a whole new realm of competitive advantage. As an evolving and maturing field, data mining is often associated with a number of myths, including those listed in Table 4.6 (Delen, 2014; Zaima, 2003).

Data mining visionaries have gained enormous competitive advantage by understanding that these myths are just that: myths.

Although the value proposition and therefore its necessity are obvious to anyone, those who carry out data mining projects (from novice to seasoned data scientist) sometimes make mistakes that result in projects with less-than-desirable outcomes. The following 16 data mining mistakes (also called *blunders, pitfalls*, or *bloopers*) are often made in practice (Nisbet et al., 2009; Shultz, 2004; Skalak, 2001), and data scientists should be aware of them and, to the extent that is possible, do their best to avoid them:

1. Selecting the wrong problem for data mining. Not every business problem can be solved with data mining (i.e., the magic bullet syndrome). When there are no representative data (large and feature rich), there cannot be a practicable data mining project.
2. Ignoring what your sponsor thinks data mining is and what it really can and cannot do. Expectation management is the key for successful data mining projects.

TABLE 4.6 Data Mining Myths

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for mainstream business applications.	The current state of the art is ready for almost any business type and/or size.
Data mining requires a separate, dedicated database.	Because of the advances in database technology, a dedicated database is not required.
Only those with advanced degrees can do data mining.	Newer Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, any company can use data mining.

3. Beginning without the end in mind. Although data mining is a process of knowledge discovery, one should have a goal/objective (a stated business problem) in mind to succeed. Because, as the saying goes, "If you don't know where you are going, you will never get there."
4. Defining the project around a foundation that your data cannot support. Data mining is all about data; that is, the biggest constraint that you have in a data mining project is the richness of the data. Knowing what the limitations of data are helps you craft feasible projects that deliver results and meet expectations.
5. Leaving insufficient time for data preparation. It takes more effort than is generally understood. The common knowledge suggests that up to one-third of the total project time is spent on data acquisition, understanding, and preparation tasks. To succeed, avoid proceeding into modeling until after your data are properly processed (aggregated, cleaned, and transformed).
6. Looking only at aggregated results, not at individual records. Data mining is at its best when the data are at a granular representation. Try to avoid unnecessarily aggregating and overly simplifying data to help data mining algorithms—they don't really need your help; they are more than capable of figuring it out themselves.
7. Being sloppy about keeping track of the data mining procedure and results. Because data mining is a discovery process that involves many iterations and experimentations, its user is highly likely to lose track of the findings. Success requires a systematic and orderly planning, execution, and tracking/recording of all data mining tasks.
8. Using data from the future to predict the future. Because of the lack of description and understanding of the data, oftentimes analysts include variables that are unknown at the time when the prediction is supposed to be made. By doing so, their prediction models produce unbelievably accurate results (a phenomenon that is often called *fool's gold*). If your prediction results are too good to be true, they usually are; in that case, the first thing that you need to look for is the incorrect use of a variable from the future.
9. Ignoring suspicious findings and quickly moving on. The unexpected findings are often the indicators of real novelties in data mining projects. Proper investigation of such oddities can lead to surprisingly pleasing discoveries.
10. Starting with a high-profile complex project that will make you a superstar. Data mining projects often fail if they are not thought out carefully from start to end. Success often comes with a systematic and orderly progression of projects from smaller/simpler to larger/complex ones. The goal should be to show incremental and continuous value added as opposed to taking on a large project that will consume resources without producing any valuable outcomes.
11. Running data mining algorithms repeatedly and blindly. Although today's data mining tools are capable of consuming data and setting algorithmic parameters to produce results, one should know how to transform the data and set the proper parameter values to obtain the best possible results. Each algorithm has its own unique way to process data, and knowing that is necessary to get the most out of each model type.
12. Ignore the subject matter experts. Understanding the problem domain and the related data requires a highly involved collaboration between the data mining and the domain experts. Working together helps the data mining expert to go beyond the syntactic representation and to obtain semantic nature (i.e., the true meaning of the variables) of the data.
13. Believing everything you are told about the data. Although it is necessary to talk to domain experts to better understand the data and the business problem, the data scientist should not take anything for granted. Validation and verification through a critical analysis is the key to intimate understanding and processing of the data.
14. Assuming that the keepers of the data will be fully on board with cooperation. Many data mining projects fail because the data mining expert did not know/understand the organizational politics. One of the biggest obstacles in data mining projects can

be the people who own and control the data. Understanding and managing the politics is a key to identify, access, and properly understand the data to produce a successful data mining project.

15. Measuring your results differently from the way your sponsor measures them. The results should talk/appeal to the end user (manager/decision maker) who will be using them. Therefore, producing the results in a measure and format that appeals to the end user tremendously increases the likelihood of true understanding and proper use of the data mining outcomes.
16. Follow the advice in a well-known quote: "If you build it, they will come": don't worry about how to serve it up. Usually, data mining experts think they have finished once they build models that meet and hopefully exceed the needs/wants/expectations of the end user (i.e., the customer). Without a proper deployment, the value deliverance of data mining outcomes is rather limited. Therefore, deployment is a necessary last step in the data mining process in which models are integrated into the organizational decision support infrastructure for enablement of better and faster decision making.

SECTION 4.7 REVIEW QUESTIONS

1. What are the privacy issues in data mining?
2. How do you think the discussion between privacy and data mining will progress? Why?
3. What are the most common myths about data mining?
4. What do you think are the reasons for these myths about data mining?
5. What are the most common data mining mistakes/blunders? How can they be alleviated or completely eliminated?

Chapter Highlights

- Data mining is the process of discovering new knowledge from databases.
- Data mining can use simple flat files as data sources, or it can be performed on data in data warehouses.
- There are many alternative names and definitions for data mining.
- Data mining is at the intersection of many disciplines, including statistics, artificial intelligence, and mathematical modeling.
- Companies use data mining to better understand their customers and optimize their operations.
- Data mining applications can be found in virtually every area of business and government, including healthcare, finance, marketing, and homeland security.
- Three broad categories of data mining tasks are prediction (classification or regression), clustering, and association.
- Similar to other IS initiatives, a data mining project must follow a systematic project management process to be successful.
- Several data mining processes have been proposed: CRISP-DM, SEMMA, KDD, for example.
- CRISP-DM provides a systematic and orderly way to conduct data mining projects.
- The earlier steps in data mining projects (i.e., understanding the domain and the relevant data) consume most of the total project time (often more than 80% of the total time).
- Data preprocessing is essential to any successful data mining study. Good data lead to good information; good information leads to good decisions.
- Data preprocessing includes four main steps: data consolidation, data cleaning, data transformation, and data reduction.
- Classification methods learn from previous examples containing inputs and the resulting class labels, and once properly trained, they are able to classify future cases.
- Clustering partitions pattern records into natural segments or clusters. Each segment's members share similar characteristics.
- A number of different algorithms are commonly used for classification. Commercial implementations include ID3, C4.5, C5, CART, CHAID, and SPRINT.
- Decision trees partition data by branching along different attributes so that each leaf node has all the patterns of one class.

- The Gini index and information gain (entropy) are two popular ways to determine branching choices in a decision tree.
- The Gini index measures the purity of a sample. If everything in a sample belongs to one class, the Gini index value is zero.
- Several assessment techniques can measure the prediction accuracy of classification models, including simple split, *k*-fold cross-validation, bootstrapping, and the area under the ROC curve.
- There are a number of methods to assess the variable importance of data mining models. Some of these methods are model type specific, some are model type agnostic.
- Cluster algorithms are used when data records do not have predefined class identifiers (i.e., it is not known to what class a particular record belongs).
- Cluster algorithms compute measures of similarity in order to group similar cases into clusters.
- The most commonly used similarity measure in cluster analysis is a distance measure.
- The most commonly used clustering algorithms are *k*-means and self-organizing maps.
- Association rule mining is used to discover two or more items (or events or concepts) that go together.
- Association rule mining is commonly referred to as market-basket analysis.
- The most commonly used association algorithm is Apriori by which frequent itemsets are identified through a bottom-up approach.
- Association rules are assessed based on their support and confidence measures.
- Many commercial and free data mining tools are available.
- The most popular commercial data mining tools are IBM SPSS Modeler and SAS Enterprise Miner.
- The most popular free data mining tools are KNIME, RapidMiner, and Weka.

Key Terms

Apriori algorithm
area under the ROC curve
association
bootstrapping
categorical data
classification
clustering
confidence
CRISP-DM
data mining
decision tree
distance measure
ensemble
entropy
Gini index
information gain
interval data
k-fold cross-validation
KNIME
knowledge discovery in databases (KDD)
lift
link analysis
Microsoft Enterprise Consortium
Microsoft SQL Server
nominal data
numeric data
ordinal data
prediction
RapidMiner
regression
SEMMA
sensitivity analysis
sequence mining
simple split
support
Weka

Questions for Discussion

1. Define *data mining*. Why are there many names and definitions for data mining?
2. What are the main reasons for the recent popularity of data mining?
3. Discuss what an organization should consider before making a decision to purchase data mining software.
4. Distinguish data mining from other analytical tools and techniques.
5. Discuss the main data mining methods. What are the fundamental differences among them?
6. What are the main data mining application areas? Discuss the commonalities of these areas that make them a prospect for data mining studies.
7. Why do we need a standardized data mining process? What are the most commonly used data mining processes?
8. Discuss the differences between the two most commonly used data mining processes.
9. Are data mining processes a mere sequential set of activities? Explain.
10. Why do we need data preprocessing? What are the main tasks and relevant techniques used in data preprocessing?
11. Discuss the reasoning behind the assessment of classification models.
12. What is the main difference between classification and clustering? Explain using concrete examples.
13. Moving beyond the chapter discussion, where else can association be used?
14. What are the privacy issues with data mining? Do you think they are substantiated?
15. What are the most common myths and mistakes about data mining?

Exercises

Teradata University Network (TUN) and Other Hands-On Exercises

1. Visit **teradatauniversitynetwork.com**. Identify case studies and white papers about data mining. Describe recent developments in the field of data mining and predictive modeling.
2. Go to **teradatauniversitynetwork.com**. Locate Web seminars related to data mining. In particular, locate and watch a seminar given by C. Imhoff and T. Zouqes. Then answer the following questions:
   a. What are some of the interesting applications of data mining?
   b. What types of payoffs and costs can organizations expect from data mining initiatives?
3. For this exercise, your goal is to build a model to identify inputs or predictors that differentiate risky customers from others (based on patterns pertaining to previous customers) and then use those inputs to predict new risky customers. This sample case is typical for this domain.

   The sample data to be used in this exercise are in Online File W4.1 in the file CreditRisk.xlsx. The data set has 425 cases and 15 variables pertaining to past and current customers who have borrowed from a bank for various reasons. The data set contains customer-related information such as financial standing, reason for the loan, employment, demographic information, and the outcome or dependent variable for credit standing, classifying each case as good or bad based on the institution's past experience.

   Take 400 of the cases as training cases and set aside the other 25 for testing. Build a decision tree model to learn the characteristics of the problem. Test its performance on the other 25 cases. Report on your model's learning and testing performance. Prepare a report that identifies the decision tree model and training parameters as well as the resulting performance on the test set. Use any decision tree software. (This exercise is courtesy of StatSoft, Inc., based on a German data set from **ftp.ics.uc,i.edu/pub/machine-learning-databases/statlog/german** renamed CreditRisk and altered.)
4. For this exercise, you will replicate (on a smaller scale) the box-office prediction modeling explained in Application Case 4.6. Download the training data set from Online File W4.2, MovieTrain.xlsx, which is in Microsoft Excel format. Use the data description given in Application Case 4.6 to understand the domain and the problem you are trying to solve. Pick and choose your independent variables. Develop at least three classification models (e.g., decision tree, logistic regression, neural networks). Compare the accuracy results using 10-fold cross-validation and percentage split techniques, use confusion matrices, and comment on the outcome. Test the models you have developed on the test set (see Online File W4.3, MovieTest.xlsx). Analyze the results with different models, and find the best classification model, supporting it with your results.
5. This exercise introduces you to association rule mining. The Excel data set baskets1ntrans.xlsx has around 2,800 observations/records of supermarket transaction products data. Each record contains the customer's ID and products that they have purchased. Use this data set to understand the relationships among products (i.e., which products are purchased together). Look for interesting relationships and add screenshots of any subtle association patterns that you might find. More specifically, answer the following questions.
   - Which association rules do you think are most important?
   - Based on some of the association rules you found, make at least three business recommendations that might be beneficial to the company. These recommendations can include ideas about shelf organization, up-selling, or cross-selling products. (Bonus points will be given to new/innovative ideas.)
   - What are the Support, Confidence, and Lift values for the following rule?

   Wine, Canned Veg → Frozen Meal
6. In this assignment, you will use a free/open source data mining tool, KNIME (**knime.org**), to build predictive models for a relatively small Customer Churn Analysis data set. You are to analyze the given data set (about the customer retention/attrition behavior for 1,000 customers) to develop and compare at least three prediction (i.e., classification) models. For example, you can include decision trees, neural networks, SVM, *k* nearest neighbor, and/or logistic regression models in your comparison. Here are the specifics for this assignment:
   - Install and use the KNIME software tool from (**knime.org**).
   - You can also use MS Excel to preprocess the data (if you need to/want to).
   - Download CustomerChurnData.csv data file from the book's Web site.
   - The data are given in comma-separated value (CSV) format. This format is the most common flat-file format that many software tools can easily open/handle (including KNIME and MS Excel).
   - Present your results in a well-organized professional document.
   - Include a cover page (with proper information about you and the assignment).
   - Make sure to nicely integrate figures (graphs, charts, tables, screenshots) within your textual description in a professional manner. The report should have six main sections (resembling CRISP-DM phases).
   - Try not to exceed 15 pages in total, including the cover (use 12-point Times New Roman fonts, and 1.5-line spacing).

Team Assignments and Role-Playing Projects

1. Examine how new data capture devices such as RFID tags help organizations accurately identify and segment their customers for activities such as targeted marketing. Many of these applications involve data mining. Scan the literature and the Web and then propose five potential new data mining applications that can use the data created with RFID technology. What issues could arise if a country's laws required such devices to be embedded in everyone's body for a national identification system?
2. Interview administrators in your college or executives in your organization to determine how data mining, data warehousing, Online Analytics Processing (OLAP), and visualization tools could assist them in their work. Write a proposal describing your findings. Include cost estimates and benefits in your report.
3. A very good repository of data that has been used to test the performance of many data mining algorithms is available at **ics.uci.edu/~mlearn/MLRepository.html**. Some of the data sets are meant to test the limits of current machine-learning algorithms and to compare their performance with new approaches to learning. However, some of the smaller data sets can be useful for exploring the functionality of any data mining software, such as RapidMiner or KNIME. Download at least one data set from this repository (e.g., Credit Screening Databases, Housing Database) and apply decision tree or clustering methods, as appropriate. Prepare a report based on your results. (Some of these exercises, especially the ones that involve large/challenging data/problem may be used as semester-long term projects.)
4. Large and feature-rich data sets are made available by the U.S. government or its subsidiaries on the Internet. For instance, see a large collection of government data sets (**data.gov**), the Centers for Disease Control and Prevention data sets (**www.cdc.gov/DataStatistics**), Surveillance, Cancer.org's Epidemiology and End Results data sets (**http://seer.cancer.gov/data**), and the Department of Transportation's Fatality Analysis Reporting System crash data sets (**www.nhtsa.gov/FARS**). These data sets are not preprocessed for data mining, which makes them a great resource to experience the complete data mining process. Another rich source for a collection of analytics data sets is listed on **KDnuggets.com** (**KDnuggets.com/datasets/index.html**).
5. Consider the following data set, which includes three attributes and a classification for admission decisions into an MBA program:

GMAT	GPA	Quantitative GMAT Score (percentile)	Decision
650	2.75	35	No
580	3.50	70	No
600	3.50	75	Yes
450	2.95	80	No
700	3.25	90	Yes
590	3.50	80	Yes
400	3.85	45	No
640	3.50	75	Yes
540	3.00	60	?
690	2.85	80	?
490	4.00	65	?

   a. Using the data shown, develop your own manual expert rules for decision making.
   b. Use the Gini index to build a decision tree. You can use manual calculations or a spreadsheet to perform the basic calculations.
   c. Use an automated decision tree software program to build a tree for the same data.

Internet Exercises

1. Visit the AI Exploratorium at **cs.ualberta.ca/~aixplore**. Click the Decision Tree link. Read the narrative on basketball game statistics. Examine the data, and then build a decision tree. Report your impressions of its accuracy. Also explore the effects of different algorithms.
2. Survey some data mining tools and vendors. Start with **fico.com** and **egain.com**. Consult **dmreview.com**, and identify some data mining products and service providers that are not mentioned in this chapter.
3. Find recent cases of successful data mining applications. Visit the Web sites of some data mining vendors, and look for cases or success stories. Prepare a report summarizing five new case studies.
4. Go to vendor Web sites (especially those of SAS, SPSS, Cognos, Teradata, StatSoft, and Fair Isaac) and look at success stories for BI (OLAP and data mining) tools. What do the various success stories have in common? How do they differ?
5. Go to **statsoft.com** (now a Dell company). Download at least three white papers on applications. Which of these applications might have used the data/text/Web mining techniques discussed in this chapter?
6. Go to **sas.com**. Download at least three white papers on applications. Which of these applications could have used the data/text/Web mining techniques discussed in this chapter?
7. Go to **spss.com** (an IBM company). Download at least three white papers on applications. Which of these applications could have used the data/text/Web mining techniques discussed in this chapter?
8. Go to **teradata.com**. Download at least three white papers on applications. Which of these applications could have used the data/text/Web mining techniques discussed in this chapter?
9. Go to **fico.com**. Download at least three white papers on applications. Which of these applications could have used the data/text/Web mining techniques discussed in this chapter?

10. Go to **salfordsystems.com**. Download at least three white papers on applications. Which of these applications could have used the data/text/Web mining techniques discussed in this chapter?
11. Go to **rulequest.com**. Download at least three white papers on applications. Which of these applications could have used the data/text/Web mining techniques discussed in this chapter?
12. Go to **KDnuggets.com**. Explore the sections on applications as well as software. Find names of at least three additional packages for data mining and text mining.

References

Chan, P., Phan, W., Prodromidis, A., & Stolfo, S. (1999). "Distributed Data Mining in Credit Card Fraud Detection." *IEEE Intelligent Systems, 14*(6), 67–74.

CRISP-DM. (2013). "Cross-Industry Standard Process for Data Mining (CRISP-DM)." **http://crisp-dm.orgwww.the-modeling-agency.com/crisp-dm.pdf** (accessed February 2, 2013).

Davenport, T. (2006, January). "Competing on Analytics." *Harvard Business Review*, 99–107.

Delen, D. (2009). "Analysis of Cancer Data: A Data Mining Approach." *Expert Systems, 26*(1), 100–112.

Delen, D. (2014). *Real-World Data Mining: Applied Business Analytics and Decision Making*. Upper Saddle River, NJ: Pearson.

Delen, D., Cogdell, D., & Kasap, N. (2012). "A Comparative Analysis of Data Mining Methods in Predicting NCAA Bowl Outcomes." *International Journal of Forecasting, 28*, 543–552.

Delen, D., Sharda, R., & Kumar, P. (2007). "Movie Forecast Guru: A Web-Based DSS for Hollywood Managers." *Decision Support Systems, 43*(4), 1151–1170.

Delen, D., Walker, G., & Kadam, A. (2005). "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods." *Artificial Intelligence in Medicine, 34*(2), 113–127.

Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). "From Knowledge Discovery in Databases." *AI Magazine, 17*(3), 37–54.

Hoffman, T. (1998, December 7). "Banks Turn to IT to Reclaim Most Profitable Customers." *Computerworld*.

Hoffman, T. (1999, April 19). "Insurers Mine for Age-Appropriate Offering." *Computerworld*.

Kohonen, T. (1982). "Self-Organized Formation of Topologically Correct Feature Maps." *Biological Cybernetics, 43*(1), 59–69.

Nemati, H., & Barko, C. (2001). "Issues in Organizational Data Mining: A Survey of Current Practices." *Journal of Data Warehousing, 6*(1), 25–36.

Nisbet, R., Miner, G., & Elder IV, J. (2009). "Top 10 Data Mining Mistakes." *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, pp. 733–754.

Quinlan, J. (1986). "Induction of Decision Trees." *Machine Learning, 1*, 81–106.

Saltelli, A. (2002). "Making Best Use of Model Evaluations to Compute Sensitivity Indices," *Computer Physics Communications, 145*, 280–297.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity Analysis in Practice – A Guide to Assessing Scientific Models*. Hoboken, NJ: John Wiley.

SEMMA. (2009). "SAS's Data Mining Process: Sample, Explore, Modify, Model, Assess." **sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html** (accessed August 2009).

Sharda, R., & Delen, D. (2006). "Predicting Box-Office Success of Motion Pictures with Neural Networks." *Expert Systems with Applications, 30*, 243–254.

Shultz, R. (2004, December 7). "Live from NCDM: Tales of Database Buffoonery." **directmag.com/news/ncdm-12-07-04/index.html** (accessed April 2009).

Skalak, D. (2001). "Data Mining Blunders Exposed!" *DB2 Magazine, 6*(2), 10–13.

Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2009). "Toward Breast Cancer Survivability Prediction Models Through Improving Training Space." *Expert Systems with Applications, 36*(10), 12200–12209.

Wald, M. (2004, February 21). "U.S. Calls Release of JetBlue Data Improper." *The New York Times*.

Wright, C. (2012). "Statistical Predictors of March Madness: An Examination of the NCAA Men's Basketball Championship." **http://economics-files.pomona.edu/GarySmith/Econ190/Wright%20March%20Madness%20Final%20Paper.pdf** (accessed February 2, 2013).

Zaima, A. (2003). "The Five Myths of Data Mining." *What Works: Best Practices in Business Intelligence and Data Warehousing,* Vol. 15. Chatsworth, CA: The Data Warehousing Institute, pp. 42–43.

Zolbanin, H., Delen, D., & Zadeh, A. (2015). "Predicting Overall Survivability in Comorbidity of Cancers: A Data Mining Approach." *Decision Support Systems*, 74, 150–161.

CHAPTER 5

Machine-Learning Techniques for Predictive Analytics

LEARNING OBJECTIVES

- Understand the basic concepts and definitions of artificial neural networks (ANN)
- Learn the different types of ANN architectures
- Understand the concept and structure of support vector machines (SVM)
- Learn the advantages and disadvantages of SVM compared to ANN
- Understand the concept and formulation of *k*-nearest neighbor (*k*NN) algorithm
- Learn the advantages and disadvantages of *k*NN compared to ANN and SVM
- Understand the basic principles of Bayesian learning and Naïve Bayes algorithm
- Learn the basics of Bayesian Belief Networks and how they are used in predictive analytics
- Understand different types of ensemble models and their pros and cons in predictive analytics

Predictive modeling is perhaps the most commonly practiced branch in data science and business analytics. It allows decision makers to estimate what the future holds by means of learning from the past (i.e., historical data). In this chapter, we study the internal structures, capabilities/limitations, and applications of the most popular predictive modeling techniques, such as artificial neural networks, support vector machines, *k*-nearest neighbor, Bayesian learning, and ensemble models. Most of these techniques are capable of addressing both classification- and regression-type prediction problems. Often, they are applied to complex prediction problems where other, more traditional techniques are not capable of producing satisfactory results. In addition to the ones covered in this chapter, other notable prediction modeling techniques include regression (linear or nonlinear), logistic regression (for classification-type prediction problems), and different types of decision trees (covered in Chapter 4).

5.1 OPENING VIGNETTE: Predictive Modeling Helps Better Understand and Manage Complex Medical Procedures

Healthcare has become one of the most important issues to have a direct impact on the quality of life in the United States and around the world. While the demand for healthcare services is increasing because of the aging population, the supply side is having problems keeping up with the level and quality of service. To close the gap, healthcare systems ought to significantly improve their operational effectiveness and efficiency. Effectiveness (doing the right thing, such as diagnosing and treating accurately) and efficiency (doing it the right way, such as using the least amount of resources and time) are the two fundamental pillars upon which the healthcare system can be revived. A promising way to improve healthcare is to take advantage of predictive modeling techniques along with large and feature-rich data sources (true reflections of medical and healthcare experiences) to support accurate and timely decision making.

According to the American Heart Association, cardiovascular disease (CVD) is the underlying cause for over 20 percent of deaths in the United States. Since 1900, CVD has been the number-one killer every year except 1918, which was the year of the great flu pandemic. CVD kills more people than the next four leading causes of deaths combined: cancer, chronic lower respiratory disease, accidents, and diabetes mellitus. Of all CVD deaths, more than half are attributed to coronary diseases. Not only does CVD take a huge toll on the personal health and well-being of the population, but also it is a great drain on the healthcare resources in the United States and elsewhere in the world. The direct and indirect costs associated with CVD for a year are estimated to be in excess of $500 billion. A common surgical procedure to cure a large variant of CVD is called *coronary artery bypass grafting* (CABG). Even though the cost of a CABG surgery depends on the patient and service provider–related factors, the average rate is between $50,000 and $100,000 in the United States. As an illustrative example, Delen et al. (2012) carried out an analytics study using various predictive modeling methods to predict the outcome of a CABG and applied an information fusion–based sensitivity analysis on the trained models to better understand the importance of the prognostic factors. The main goal was to illustrate that predictive and explanatory analysis of large and feature-rich data sets provides invaluable information to make more efficient and effective decisions in healthcare.

THE RESEARCH METHOD

Figure 5.1 shows the model development and testing process used by Delen et al. (2012). They employed four different types of prediction models (artificial neural networks, support vector machines, and two types of decision trees—C5 and CART) and went through

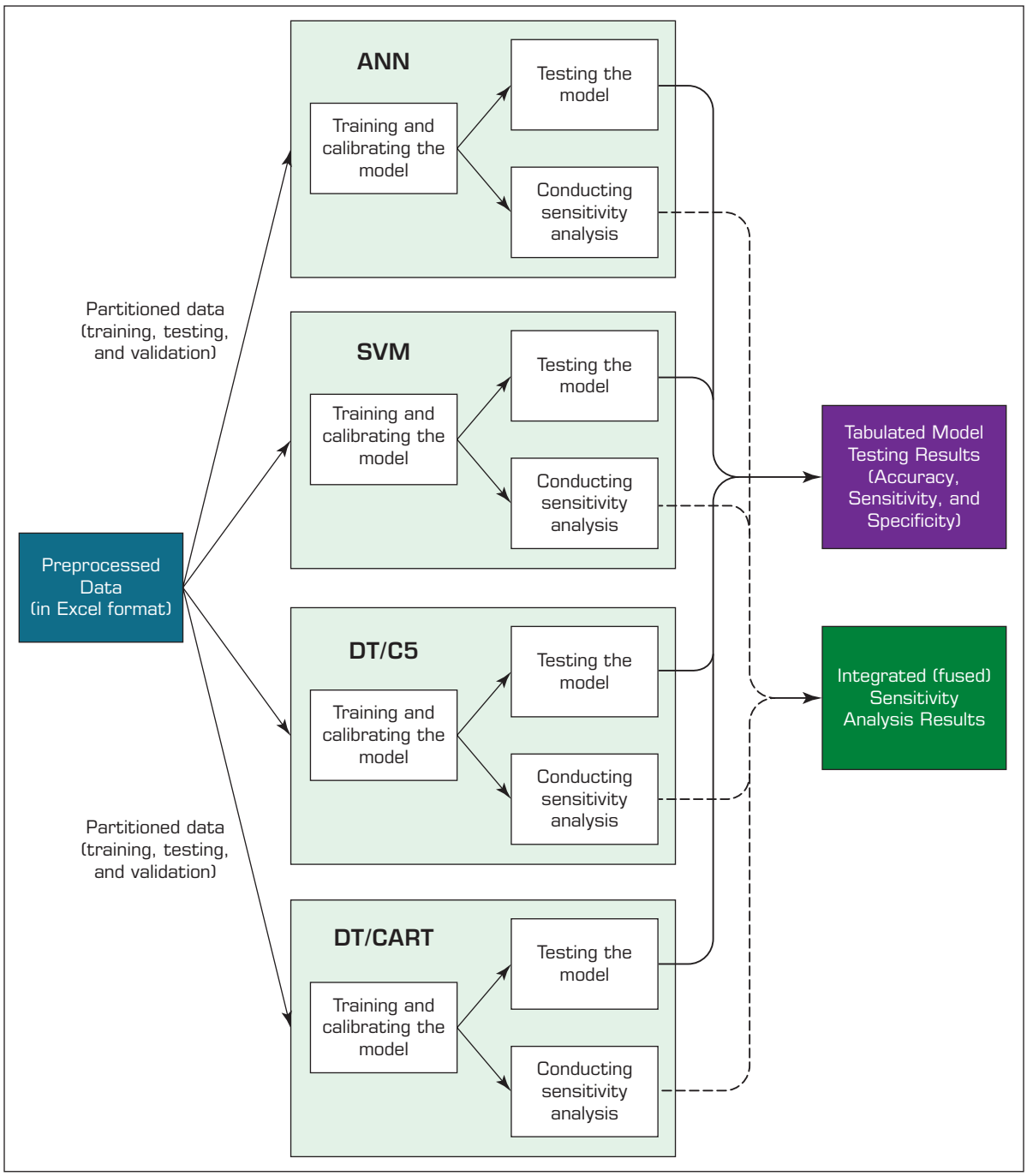


FIGURE 5.1 **Process Map for Training and Testing of the Four Predictive Models.**

a large number of experimental runs to calibrate the modeling parameters for each model type. Once the models were developed, the researchers went to the text data set. Finally, the trained models were exposed to a sensitivity analysis procedure that measured the contribution of the variables. Table 5.1 shows the test results for the four different types of prediction models.

TABLE 5.1 Prediction Accuracy Results for All Four Model Types Based on the Test Data Set

Model Type[1]		Confusion Matrices[2] Pos (1)	Neg (0)	Accuracy[3]	Sensitivity[3]	Specificity[3]
ANN	Pos (1)	749	230	74.72%	76.51%	72.93%
	Neg (0)	265	714			
SVM	Pos (1)	876	103	87.74%	89.48%	86.01%
	Neg (0)	137	842			
C5	Pos (1)	876	103	79.62%	80.29%	78.96%
	Neg (0)	137	842			
CART	Pos (1)	660	319	71.15%	67.42%	74.87%
	Neg (0)	246	733			

[1]Acronyms for model types: artificial neural networks (ANN), support vector machines (SVM), popular decision tree algorithm (C5), classification and regression trees (CART).

[2]Prediction results for the test data samples are shown in a confusion matrix where the rows represent the actuals and columns represent the predicted cases.

[3]Accuracy, sensitivity, and specificity are the three performance measures that were used in comparing the four prediction models.

THE RESULTS

In this study, Delen et al. (2012) showed the power of data mining in predicting the outcome and in analyzing the prognostic factors of complex medical procedures such as CABG surgery. The researchers showed that using a number of prediction methods (as opposed to only one) in a competitive experimental setting has the potential to produce better predictive as well as explanatory results. Among the four methods that they used, SVM produced the best results with prediction accuracy of 88 percent on the test data sample. The information fusion–based sensitivity analysis revealed the ranked importance of the independent variables. Some of the top variables identified in this analysis having to overlap with the most important variables identified in previously conducted clinical and biological studies confirm the validity and effectiveness of the proposed data mining methodology.

From the managerial standpoint, clinical decision support systems that use the outcome of data mining studies (such as the ones presented in this case study) are not meant to replace healthcare managers and/or medical professionals. Rather, they intend to support them in making accurate and timely decisions to optimally allocate resources to increase the quantity and quality of medical services. There still is a long way to go before we can see these decision aids being used extensively in healthcare practices. Among others, there are behavioral, ethical, and political reasons for this resistance to adoption. Maybe the need and government incentives for better healthcare systems will expedite the adoption.

QUESTIONS FOR THE OPENING VIGNETTE

1. Why is it important to study medical procedures? What is the value in predicting outcomes?
2. What factors do you think are the most important in better understanding and managing healthcare? Consider both managerial and clinical aspects of healthcare.
3. What would be the impact of predictive modeling on healthcare and medicine? Can predictive modeling replace medical or managerial personnel?

4. What were the outcomes of the study? Who can use these results? How can the results be implemented?
5. Search the Internet to locate two additional cases that used predictive modeling to understand and manage complex medical procedures.

WHAT WE CAN LEARN FROM THIS VIGNETTE

As you will see in this chapter, predictive modeling techniques can be applied to a wide range of problem areas, from standard business problems of assessing customer needs to understanding and enhancing the efficiency of production processes to improving healthcare and medicine. This vignette illustrates an innovative application of predictive modeling to better predict, understand, and manage CABG procedures. As the results indicate, these sophisticated analytics techniques are capable of predicting and explaining such complex phenomena. *Evidence-based medicine* is a relatively new term coined in the healthcare arena where the main idea is to dig deeply into past experiences to discover new and useful knowledge to improve medical and managerial procedures in healthcare. As we all know, healthcare needs all the help that it can get. Compared to traditional research, which is clinical and biological in nature, data-driven studies provide an out-of-the-box view to medicine and management of medical systems.

Sources: D. Delen, A. Oztekin, and L. Tomak, "An Analytic Approach to Better Understanding and Management of Coronary Surgeries," *Decision Support Systems,* Vol. 52, No. 3, 2012, pp. 698–705; and American Heart Association, "Heart Disease and Stroke Statistics," **heart.org** (accessed May 2018).

5.2 BASIC CONCEPTS OF NEURAL NETWORKS

Neural networks represent a brain metaphor for information processing. These models are biologically inspired rather than an exact replica of how the brain actually functions. Neural networks have been shown to be very promising systems in many forecasting and business classification applications due to their ability to "learn" from the data, their nonparametric nature (i.e., no rigid assumptions), and their ability to generalize. **Neural computing** refers to a pattern-recognition methodology for machine learning. The resulting model from neural computing is often called an **artificial neural network (ANN)** or a **neural network**. Neural networks have been used in many business applications for **pattern recognition**, forecasting, prediction, and classification. Neural network computing is a key component of any data science and business analytics toolkit. Applications of neural networks abound in finance, marketing, manufacturing, operations, information systems, and so on.

Because we cover neural networks, especially the feed-forward, multi-layer, perception-type prediction modeling–specific neural network architecture, in Chapter 6 (which is dedicated to deep learning and cognitive computing) as a primer to understanding deep learning and deep neural networks, in this section, we provide only a brief introduction to the vast variety of neural network models, methods, and applications.

The human brain possesses bewildering capabilities for information processing and problem solving that modern computers cannot compete with in many aspects. It has been postulated that a model or a system that is enlightened and supported by results from brain research and has a structure similar to that of biological neural networks could exhibit similar intelligent functionality. Based on this bottom-up approach,

ANN (also known as *connectionist models, parallel distributed processing models, neuromorphic systems,* or simply *neural networks*) has been developed as biologically inspired and plausible models for various tasks.

Biological neural networks are composed of many massively interconnected **neurons**. Each neuron possesses **axons** and **dendrites**, fingerlike projections that enable a neuron to communicate with its neighboring neurons by transmitting and receiving electrical and chemical signals. More or less resembling the structure of their biological counterparts, ANN are composed of interconnected, simple processing elements called *artificial neurons*. When processing information, the processing elements in ANN operate concurrently and collectively, similar to biological neurons. ANN possess some desirable traits similar to those of biological neural networks, such as the abilities to learn, self-organize, and support fault tolerance.

Coming along a winding journey, ANN have been investigated by researchers for more than half a century. The formal study of ANN began with the pioneering work of McCulloch and Pitts in 1943. Inspired by the results of biological experiments and observations, McCulloch and Pitts (1943) introduced a simple model of a binary artificial neuron that captured some of the functions of biological neurons. Using information-processing machines to model the brain, McCulloch and Pitts built their neural network model using a large number of interconnected artificial binary neurons. From these beginnings, neural network research became quite popular in the late 1950s and early 1960s. After a thorough analysis of an early neural network model (called the **perceptron**, which used no hidden layer) as well as a pessimistic evaluation of the research potential by Minsky and Papert in 1969, interest in neural networks diminished.

During the past two decades, there has been an exciting resurgence in ANN studies due to the introduction of new network topologies, new activation functions, and new learning algorithms as well as progress in neuroscience and cognitive science. Advances in theory and methodology have overcome many of the obstacles that hindered neural network research a few decades ago. Evidenced by the appealing results of numerous studies, neural networks are gaining in acceptance and popularity. In addition, the desirable features in neural information processing make neural networks attractive for solving complex problems. ANN have been applied to numerous complex problems in a variety of application settings. The successful use of neural network applications has inspired renewed interest from industry and business. With the emergence of deep neural networks (as part of the rather recent deep learning phenomenon), the popularity of neural networks (with a "deeper" architectural representation and much-enhanced analytics capabilities) hit an unprecedented high, creating mile-high expectations from this new generation of neural networks. Deep neural networks are covered in detail in Chapter 6.

Biological versus Artificial Neural Networks

The human brain is composed of special cells called *neurons*. These cells do not die and replenish when a person is injured (all other cells reproduce to replace themselves and then die). This phenomenon might explain why humans retain information for an extended period of time and start to lose it when they get old—as the brain cells gradually start to die. Information storage spans sets of neurons. The brain has anywhere from 50 billion to 150 billion neurons of which there are more than 100 different kinds. Neurons are partitioned into groups called *networks*. Each network contains several thousand highly interconnected neurons. Thus, the brain can be viewed as a collection of neural networks.

The ability to learn and to react to changes in our environment requires intelligence. The brain and the central nervous system control thinking and intelligent behavior. People who suffer brain damage have difficulty learning and reacting to changing environments. Even so, undamaged parts of the brain can often compensate with new learning.

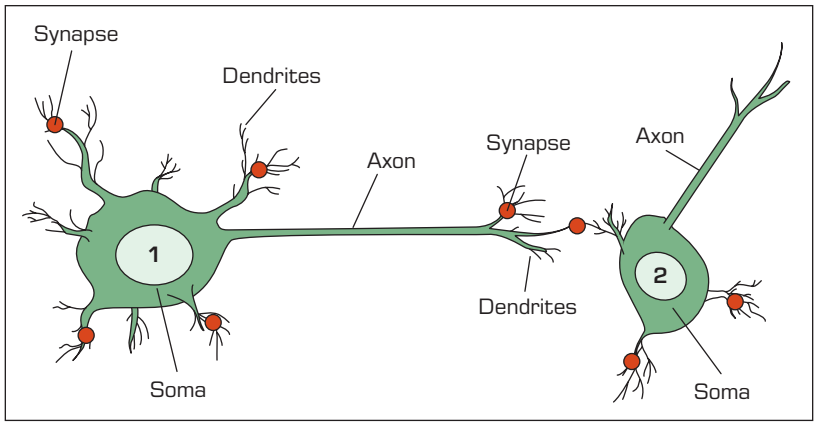


FIGURE 5.2 **Portion of a Biological Neural Network: Two Interconnected Cells/Neurons.**

A portion of a network composed of two cells is shown in Figure 5.2. The cell itself includes a **nucleus** (the central processing portion of the neuron). To the left of cell 1, the dendrites provide input signals to the cell. To the right, the axon sends output signals to cell 2 via the axon terminals. These axon terminals merge with the dendrites of cell 2. Signals can be transmitted unchanged, or they can be altered by synapses. A **synapse** is able to increase or decrease the strength of the connection between neurons and cause excitation or inhibition of a subsequent neuron. This is how information is stored in the neural networks.

An ANN emulates a biological neural network. Neural computing actually uses a very limited set of concepts from biological neural systems (see Technology Insights 5.1). It is more of an analogy to the human brain than an accurate model of it. Neural concepts usually are implemented as software simulations of the massively parallel processes involved in processing interconnected elements (also called artificial neurons, or *neurodes*) in a network architecture. The artificial neuron receives inputs analogous to the electrochemical impulses that dendrites of biological neurons receive from other neurons. The output of the artificial neuron corresponds to signals sent from a biological neuron over its axon. These artificial signals can be changed by weights in a manner similar to the physical changes that occur in the synapses (see Figure 5.3).

Several ANN paradigms have been proposed for applications in a variety of problem domains. Perhaps the easiest way to differentiate among the various neural models is on the basis of the way they structurally emulate the human brain, process information, and learn to perform their designated tasks.

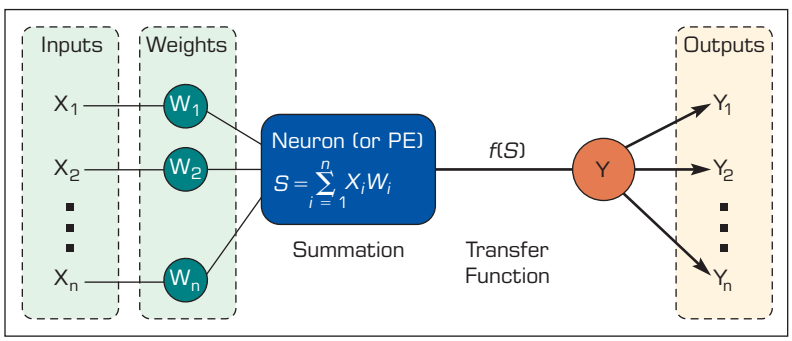


FIGURE 5.3 **Processing Information in an Artificial Neuron.**

TECHNOLOGY INSIGHTS 5.1 The Relationship between Biological and Artificial Neural Networks

The following list shows some of the relationships between biological and artificial networks.

Biological	Artificial
Soma	Node
Dendrites	Input
Axon	Output
Synapse	Weight
Slow	Fast
Many neurons (10^9)	Few neurons (a dozen to hundreds of thousands)

Sources: L. Medsker and J. Liebowitz, *Design and Development of Expert Systems and Neural Networks,* Macmillan, New York, 1994, p. 163; and F. Zahedi, *Intelligent Systems for Business: Expert Systems with Neural Networks,* Wadsworth, Belmont, CA, 1993.

Because they are biologically inspired, the main processing elements of a neural network are individual neurons, analogous to the brain's neurons. These artificial neurons receive the information from other neurons or external input stimuli, perform a transformation on the inputs, and then pass on the transformed information to other neurons or external outputs. This is similar to how it is currently thought that the human brain works. Passing information from neuron to neuron can be thought of as a way to activate, or trigger, a response from certain neurons based on the information or stimulus received.

How information is processed by a neural network is inherently a function of its structure. Neural networks can have one or more layers of neurons. These neurons can be highly or fully interconnected, or only certain layers can be connected. Connections between neurons have an associated weight. In essence, the "knowledge" possessed by the network is encapsulated in these interconnection weights. Each neuron calculates a weighted sum of the incoming neuron values, transforms this input, and passes on its neural value as the input to subsequent neurons. Typically, although not always, this input/output transformation process at the individual neuron level is performed in a nonlinear fashion.

Application Case 5.1 provides an interesting example of the use of neural networks as a prediction tool in the mining industry.

Application Case 5.1 Neural Networks Are Helping to Save Lives in the Mining Industry

In the mining industry, most of the underground injuries and fatalities are due to rock falls (i.e., fall of hanging wall/roof). The method that has been used for many years in the mines when determining the integrity of the hanging wall is to tap the hanging wall with a sounding bar and listen to the sound emitted. An experienced miner can differentiate an intact/solid hanging wall from a detached/loose hanging wall by the sound that is emitted. This method is subjective. The Council for Scientific and Industrial Research (CSIR) in South Africa has developed a device that assists any miner in making an objective decision when determining the integrity of the hanging wall. A trained neural network model is embedded into the device. The device then records the sound emitted when a hanging wall is tapped. The sound is then preprocessed before being input into a trained neural network model, which classifies the hanging wall as either intact or detached.

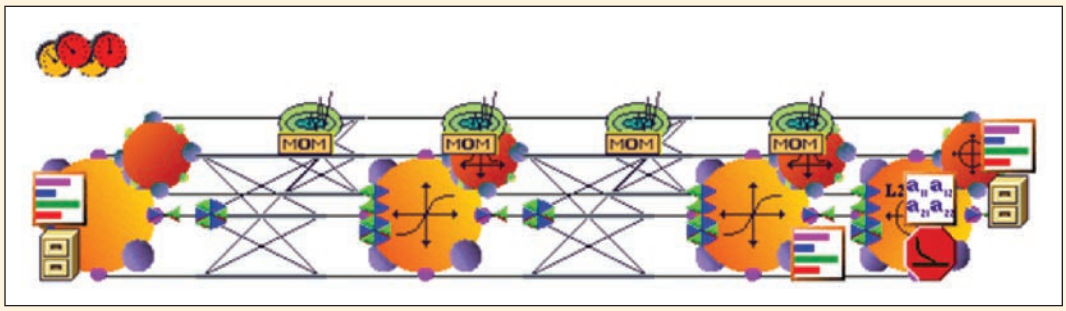
Teboho Nyareli, who holds a master's degree in electronic engineering from the University of Cape Town in South Africa and works as a research engineer at CSIR, used NeuroSolutions, a popular artificial neural network modeling software

developed by NeuroDimensions, Inc., to develop the classification-type prediction models. The multilayer perceptron-type ANN architecture that he built achieved better than 70 percent prediction accuracy on the hold-out sample. In 2018, the prototype system was undergoing a final set of tests before it was deployed as a decision aid and then the commercialization phase followed. The following figure shows a snapshot of NeuroSolution's model building workspace, called the *breadboard*.

Source: Used with permission from NeuroSolutions, customer success story, **neurosolutions.com/resources/nyareli.html** (accessed May 2018).

Questions for Case 5.1

1. How did neural networks help save lives in the mining industry?
2. What were the challenges, the proposed solution, and the results?



SECTION 5.2 REVIEW QUESTIONS

1. What is an ANN?
2. What are the commonalities and differences between biological and artificial neural networks?
3. What types of business problems can be solved with ANN?

5.3 NEURAL NETWORK ARCHITECTURES

There are several neural network architectures designed to solve different types of problems (Haykin, 2009). The most common ones include feedforward (multilayer perceptron with backpropagation), associative memory, recurrent networks, Kohonen's self-organizing feature maps, and Hopfield networks. The feedforward multi-layer perceptron-type network architecture activates the neurons (and learns the relationship between input variables and the output variable) in one direction (from input layer to the output layer, going through one or more middle/hidden layers). This neural network architecture will be covered in detail in Chapter 6; hence, the details will be skipped in this section. In contrast to feedforward neural network architecture, Figure 5.4 shows a pictorial representation of a recurrent neural network architecture where the connections between the layers are not unidirectional; rather, there are many connections in every direction between the layers and neurons, creating a complex connection structure. Many experts believe that this multidirectional connectedness better mimics the way biological neurons are structured in the human brain.

Kohonen's Self-Organizing Feature Maps

First introduced by the Finnish professor Teuvo Kohonen, **Kohonen's self-organizing feature map** (Kohonen networks, or SOM in short) provides a way to represent multidimensional data in much lower dimensional spaces, usually one or two dimensions.

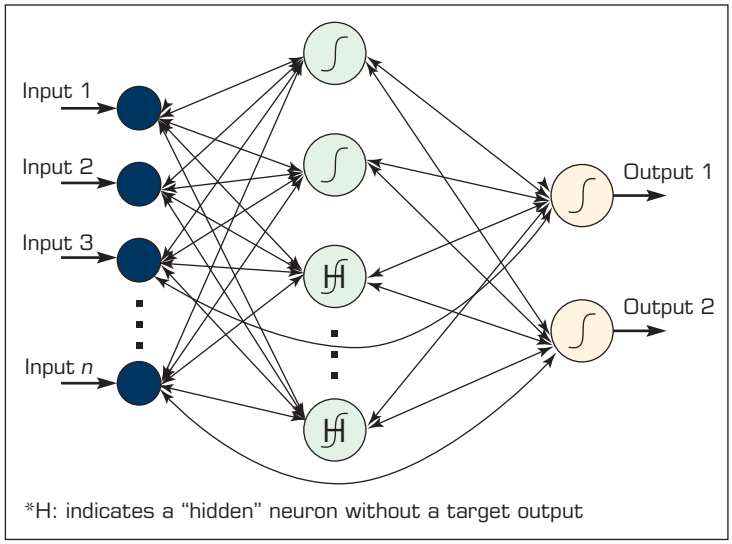


FIGURE 5.4 **Recurrent Neural Network Architecture.**

One of the most interesting aspects of SOM is that they learn to classify data without supervision (i.e., there is no output vector). Remember that in supervised learning techniques, such as backpropagation, the training data consist of vector pairs—an input vector and a target vector. Because of its self-organizing capability, SOM are commonly used for clustering tasks where a group of cases is assigned an arbitrary number of naturals groups. Figure 5.5a illustrates a very small Kohonen network of 4×4 nodes connected to the input layer (with three inputs), representing a two-dimensional vector.

Hopfield Networks

The Hopfield network is another interesting neural network architecture, first introduced by John Hopfield (1982). Hopfield demonstrated in a series of research articles in the early 1980s how highly interconnected networks of nonlinear neurons can be extremely

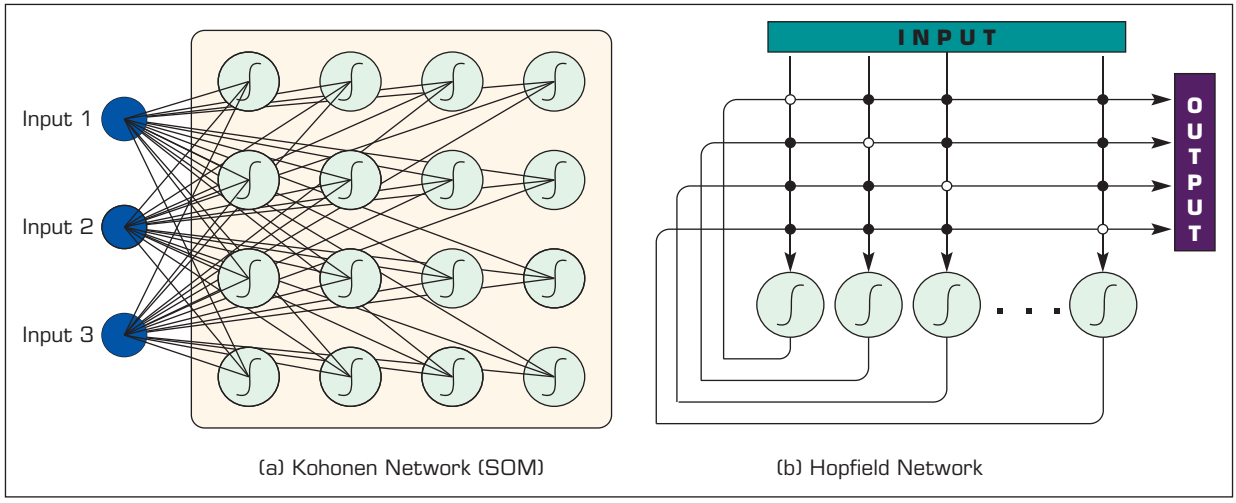


FIGURE 5.5 **Graphical Depiction of Kohonen and Hopfield ANN Structures.**

effective in solving complex computational problems. These networks were shown to provide novel and quick solutions to a family of problems stated in terms of a desired objective subject to a number of constraints (i.e., constraint optimization problems). One of the major advantages of Hopfield neural networks is the fact that their structure can be realized on an electronic circuit board, possibly on a very large-scale integration (VLSI) circuit, to be used as an online solver with a parallel-distributed process. Architecturally, a general Hopfield network is represented as a single large layer of neurons with total interconnectivity; that is, each neuron is connected to every other neuron within the network (see Figure 5.5b).

Ultimately, the architecture of a neural network model is driven by the task it is intended to carry out. For instance, neural network models have been used as classifiers, as forecasting tools, as customer segmentation mechanisms, and as general optimizers. As shown later in this chapter, neural network classifiers are typically multilayer models in which information is passed from one layer to the next, with the ultimate goal of mapping an input to the network to a specific category, as identified by an output of the network. A neural model used as an optimizer, in contrast, can be a single layer of neurons, can be highly interconnected, and can compute neuron values iteratively until the model converges to a stable state. This stable state represents an optimal solution to the problem under analysis.

Application Case 5.2 summarizes the use of predictive modeling (e.g., neural networks) in addressing emerging problems in the electric power industry.

Application Case 5.2 Predictive Modeling Is Powering the Power Generators

The electrical power industry produces and delivers electric energy (electricity or power) to both residential and business customers wherever and whenever they need it. Electricity can be generated from a multitude of sources. Most often, electricity is produced at a power station using electromechanical generators that are driven by heat engines fueled by chemical combustion (by burning coal, petroleum, or natural gas) or nuclear fusion (by a nuclear reactor). Generation of electricity can also be accomplished by other means, such as kinetic energy (through falling/flowing water or wind that activates turbines), solar energy (through the energy emitted by sun, either light or heat), or geothermal energy (through the steam or hot water coming from deep layers of the earth). Once generated, electric energy is distributed through a power grid infrastructure.

Even though some energy-generation methods are favored over others, all forms of electricity generation have positive and negative aspects. Some are environmentally favored but are economically unjustifiable; others are economically superior but environmentally prohibitive. In a market economy, the options with fewer overall costs are generally chosen above all other sources. It is not clear yet which form can best meet the necessary demand for electricity without permanently damaging the environment. Current trends indicate that increasing the shares of renewable energy and distributed generation from mixed sources has the promise of reducing/balancing environmental and economic risks.

The electrical power industry is a highly regulated, complex business endeavor. There are four distinct roles that companies choose to participate in: power producers, transmitters, distributers, and retailers. Connecting all of the producers to all of the customers is accomplished through a complex structure, called the power grid. Although all aspects of the electricity industry are witnessing stiff competition, power generators are perhaps the ones getting the lion's share of it. To be competitive, producers of power need to maximize the use of their variety of resources by making the right decisions at the right rime.

StatSoft, one of the fastest growing providers of customized analytics solutions, developed

(*Continued*)

Application Case 5.2 (Continued)

integrated decision support tools for power generators. Leveraging the data that come from the production process, these data mining–driven software tools help technicians and managers rapidly optimize the process parameters to maximize the power output while minimizing the risk of adverse effects. Following are a few examples of what these advanced analytics tools, which include ANN and SVM, can accomplish for power generators.

- ***Optimize Operation Parameters***
  Problem: A coal-burning 300 MW multicyclone unit required optimization for consistent high flame temperatures to avoid forming slag and burning excess fuel oil.
  Solution: Using StatSoft's predictive modeling tools (along with 12 months of three-minute historical data), optimized control parameter settings for stoichiometric ratios, coal flows, primary air, tertiary air, and split secondary air damper flows were identified and implemented.
  Results: After optimizing the control parameters, flame temperatures showed strong responses, resulting in cleaner combustion for higher and more stable flame temperatures.
- ***Predict Problems Before They Happen***
  Problem: A 400 MW coal-fired DRB-4Z burner required optimization for consistent and robust low NOx operations to avoid excursions and expensive downtime. Identify root causes of ammonia slip in a selective noncatalytic reduction process for NOx reduction.
  Solution: Apply predictive analytics methodologies (along with historical process data) to predict and control variability; then target processes for better performance, thereby reducing both average NOx and variability.
  Results: Optimized settings for combinations of control parameters resulted in consistently lower NOx emissions with less variability (and no excursions) over continued operations at low load, including predicting failures or unexpected maintenance issues.
- ***Reduce Emission (NOx, CO)***
  Problem: While NOx emissions for higher loads were within acceptable ranges, a 400 MW coal-fired DRB-4Z burner was not optimized for low-NOx operations under low load (50–175 MW).
  Solution: Using data-driven predictive modeling technologies with historical data, optimized parameter settings for changes to airflow were identified, resulting in a set of specific, achievable input parameter ranges that were easily implemented into the existing DCS (digital control system).
  Results: After optimization, NOx emissions under low-load operations were comparable to NOx emissions under higher loads.

As these specific examples illustrate, there are numerous opportunities for advanced analytics to make a significant contribution to the power industry. Using data and predictive models could help decision makers get the best efficiency from their production system while minimizing the impact on the environment.

Questions for Case 5.2

1. What are the key environmental concerns in the electric power industry?
2. What are the main application areas for predictive modeling in the electric power industry?
3. How was predictive modeling used to address a variety of problems in the electric power industry?

Source: Based on the StatSoft, Success Stories, **statsoft.com/Portals/0/Downloads/EPRI.pdf** (accessed June 2018) and the **statsoft.fr/pdf/QualityDigest_Dec2008.pdf** (accessed February 2018).

SECTION 5.3 REVIEW QUESTIONS

1. What are the most popular neural network architectures?
2. What types of problems are solved with Kohonen SOM ANN architecture?
3. How does Hopfield ANN architecture work? To what type of problems can it be applied?

5.4 SUPPORT VECTOR MACHINES

Support vector machines are among the popular machine-learning techniques, mostly because of their superior predictive power and their theoretical foundation. SVM are among the supervised learning techniques that produce input-output functions from a set of labeled training data. The function between the input and output vectors can be either a classification (used to assign cases into predefined classes) or a regression (used to estimate the continuous numerical value of the desired output). For classification, nonlinear kernel functions are often used to transform input data (naturally representing highly complex nonlinear relationships) to a high-dimensional feature space in which the input data become linearly separable. Then, the maximum-margin hyperplanes are constructed to optimally separate the output classes from each other in the training data.

Given a classification-type prediction problem, generally speaking, many linear classifiers (hyperplanes) can separate the data into multiple subsections, each representing one of the classes (see Figure 5.6a where the two classes are represented with circles ["○"] and squares ["■"]). However, only one hyperplane achieves the maximum separation between the classes (see Figure 5.6b where the hyperplane and the two maximum margin hyperplanes are separating the two classes).

Data used in SVM can have more than two dimensions (i.e., two distinct classes). In that case, we would be interested in separating data using the $n - 1$ dimensional hyperplane, where n is the number of dimensions (i.e., class labels). This can be seen as a typical form of linear classifier where we are interested in finding the $n - 1$ hyperplane so that the distance from the hyperplanes to the nearest data points are maximized. The assumption is that the larger the margin or distance between these parallel hyperplanes, the better the generalization power of the classifier (i.e., prediction power of the SVM model). If such hyperplanes exist, they can be mathematically represented using quadratic optimization modeling. These hyperplanes are known as the maximum-margin hyperplane, and such a linear classifier is known as a *maximum-margin classifier*.

In addition to their solid mathematical foundation in statistical learning theory, SVM have also demonstrated highly competitive performance in numerous real-world prediction problems, such as medical diagnosis, bioinformatics, face/voice recognition, demand

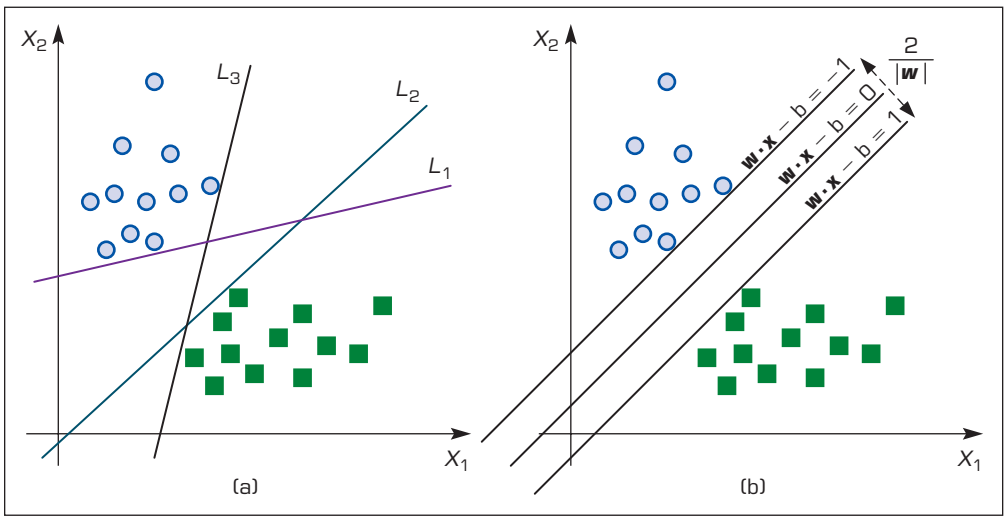


FIGURE 5.6 Separation of the Two Classes Using Hyperplanes.

forecasting, image processing, and text mining, which has established SVM as among the most popular analytics tools for knowledge discovery and data mining. Similar to artificial neural networks, SVM possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy. Therefore, SVM are of particular interest to modeling highly nonlinear, complex problems, systems, and processes. In the research study summarized in Application Case 5.3, SVM were better than other machine-learning methods in predicting and characterizing injury severity risk factors in automobile crashes.

Application Case 5.3

Identifying Injury Severity Risk Factors in Vehicle Crashes with Predictive Analytics

As technology keeps advancing, new and improved safety measures are being developed and incorporated into vehicles and roads to prevent crashes from happening and/or reduce the impact of the injury sustained by passengers caused by such incidents. Despite the extent of these efforts, the number of vehicle crashes and the resulting injuries are increasing worldwide. For instance, according to the National Highway Traffic Safety Administration (NHTSA), in the United States more than 6 million traffic accidents claim over 30,000 lives and injure more than 2 million people each year (NHTSA, 2014). The latest NHTSA report presented to the U.S. Congress in April 2014 stated that in 2012, highway fatalities in the United States reached 33,561, which is an increase of 1,082 over the previous year (Friedman, 2014). In the same year, an estimated 2.36 million people were injured in motor vehicle traffic crashes compared to 2.22 million in 2011. As a result, an average of nearly four lives were lost and nearly 270 people were injured on America's roadways every hour in 2012. In addition to the staggering number of fatalities and injuries, these traffic accidents also cost the taxpayers more than $230 billion. Hence, addressing road safety is a major problem in the United States.

Root causes of traffic accidents and crash-related injury severity are of special concern to the general public and to researchers (in academia, government, and industry) because such investigation aimed not only at prevention of crashes but also at reduction of their severe outcomes, potentially saving many lives and money. In addition to laboratory- and experimentation-based engineering research methods, another way to address the issue is to identify the most probable factors that affect injury severity by mining the historical data on vehicle crashes. Thorough understanding of the complex circumstances in which drivers and/or passengers are more likely to sustain severe injuries or even be killed in a vehicle crash can mitigate the risks involved to a great extent, thereby saving lives due to crashes. Many factors were found to have an impact on the severity of injury sustained by occupants in the event of a vehicle accident. These factors include behavioral or demographic features of the occupants (e.g., drug and/or alcohol levels, seatbelt or other restraining system usage, gender and age of the driver), crash-related situational characteristics (e.g., road surface/type/situation, direction of impact, strike versus struck, number of cars and/or other objects involved), environmental factors at the time of the accident (weather conditions, visibility and/or light conditions, time of the day, etc.), and the technical characteristics of the vehicle itself (age, weight, body type, etc.).

The main goal of this analytic study was to determine the most prevailing risk factors and their relative importance/significance in influencing the likelihood of increasing severity of injury caused by vehicle crashes. The crashes examined in this study included a collection of geographically well-represented samples. To have a consistent sample, the data set comprised only collations of specific types: single or multi-vehicle head-on collisions, single or multi-vehicle angled collisions, and single-vehicle fixed-object collisions. To obtain reliable and accurate results, this investigative study employed the most prevalent machine-learning techniques to identify the significance of crash-related factors as they relate to the changing levels of injury severity in vehicle crashes and compared the different machine-learning techniques.

The Research Method

The methodology employed in this study follows a very well-known standardized analytics process, namely cross-industry standard process for data mining (CRISP-DM). As is the case in any analytics project, a significant portion of the project time was devoted to the acquisition, integration, and preprocessing of data. Then, the preprocessed, analytics-ready data were used to build several different prediction models. Using a set of standard metrics, researchers assessed the outcomes of these models and compared them. In the final stage, sensitivity analyses were used to identify the most prevailing injury-severity related risk factors.

To effectively and efficiently perform the individual tasks in the proposed methodology, several statistical and data mining software tools were used. Specifically, JMP (a statistical and data mining software tool developed by SAS Institute), Microsoft Excel, and Tableau were used for inspecting, understanding, and preprocessing the data; IBM SPSS Modeler and KNIME were used for data merging, predictive model building, and sensitivity analysis.

The National Automotive Sampling System General Estimates System (NASS GES) data set was used for covered accidents in the years 2011 and 2012. The complete data set was obtained in the form of three separate flat/text files—accident, vehicle, and person. The accident files contained specific characteristics about road conditions, environmental conditions, and crash-related settings. The vehicle files included a large number of variables about the specific features of the vehicle involved in the crash. The person files provided detailed demographics, injury, and situational information about the occupants (i.e., driver and the passengers) impacted in the crash. To consolidate the data into a single database, the two years of data were merged within each file types (i.e., accident, person, vehicle), and the resulting files were combined using unique accident, vehicle, and person identifiers to create a single data set. After the data consolidation/aggregation, the resulting data set included person-level records—one record per person involved in a reported vehicle crash. At this point in the process (before the data cleaning, preprocessing and slicing/dicing), the complete data set included 279,470 unique records (i.e., persons/occupants involved in crashes) and more than 150 variables (a combination of accident, person, and vehicle related characteristics). Figure 5.7 graphically illustrates the individual steps involved in the processing of data.

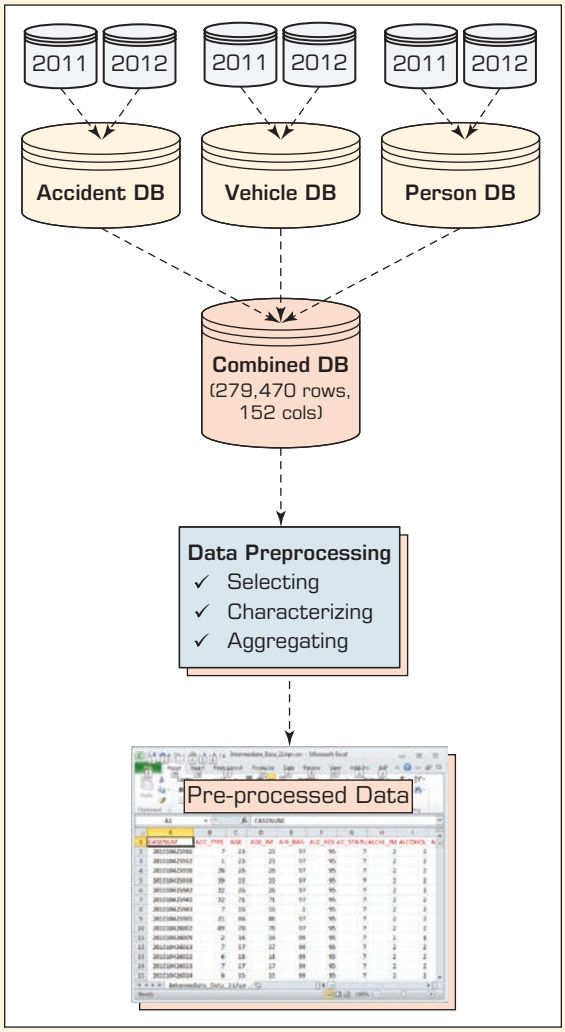


FIGURE 5.7 **Data Acquisition/Merging/Preparation Process.** *Source:* Microsoft Excel 2010, Microsoft Corporation.

Of all the variables—directly obtained from the GES databases and the ones that were derived/recalculated using the existing GES variables—29 were selected as relevant and potentially influential in determining the varying levels of injury severity involved in vehicle crashes. This extent of variables was expected to provide a rich description of the people and the vehicle involved in the accident: the specifics of the environmental conditions at the time of the crash, the settings surrounding the crash itself, and the time and place of the crash. Table 5.2 lists

(*Continued*)

Application Case 5.3 (Continued)

TABLE 5.2 List of Variables Included in the Study

Variable	Description	Data Type	Descriptive Statistics[1]	Missing (%)
AIR_BAG	Airbag deployed	Binary	Yes: 52, no: 26	5.2
ALC_RES	Alcohol test results	Numeric	12.68 (15.05)	0.4
BDYTYP_IMN	Vehicle body type	Nominal	Sedan: 34, Sm-SUV: 13	3.2
DEFORMED	Extent of damage	Nominal	Major: 43, minor: 22	3.7
DRINKING	Alcohol involvement	Binary	Yes: 4, no: 67	28.8
AGE	Age of person	Numeric	36.45 (18.49)	6.9
DRUGRES1	Drug test results	Binary	Yes: 2, no: 72	25.5
EJECT_IM	Ejection	Binary	Yes: 2, no: 93	4.9
FIRE_EXP	Fire occurred	Binary	Yes: 3, no: 97	0.0
GVWR	Vehicle weight category	Nominal	Small: 92, large: 5	2.9
HAZ_INV	Hazmat involved	Binary	Yes: 1, no: 99	0.0
HOUR_IMN	Hour of day	Nominal	Evening: 39, noon: 32	1.2
INT_HWY	Interstate highway	Binary	Yes: 13, no: 86	0.7
J_KNIFE	Jackknife	Binary	Yes: 4, no: 95	0.2
LGTCON_IM	Light conditions	Nominal	Daylight: 70, dark: 25	0.3
MANCOL_IM	Manner of collision	Nominal	Front: 34, angle: 28	0.0
MONTH	Month of year	Nominal	Oct: 10, Dec: 9	0.0
NUMINJ_IM	Number of injured	Numeric	1.23 (4.13)	0.0
PCRASH1_IMN	Precrash movement	Nominal	Going str.: 52, stopped: 14	1.3
REGION	Geographic region	Nominal	South: 42, Midwest: 24	0.0
REL_ROAD	Relation to traffic way	Nominal	Roadway: 85, median: 9	0.1
RELJCT1_IM	At a junction	Binary	Yes: 4, no: 96	0.0
REST_USE_N	Restraint system used	Nominal	Yes: 76, no: 4	7.4
SEX_IMN	Gender of driver	Binary	Male: 54, female: 43	3.1
TOWED_N	Car towed	Binary	Yes: 49, no: 51	0.0
VEH_AGE	Age of vehicle	Numeric	8.96 (4.18)	0.0
WEATHR_IM	Weather condition	Nominal	Clear: 73, cloudy: 14	0.0
WKDY_IM	Weekday	Nominal	Friday: 17, Thursday 15	0.0
WRK_ZONE	Work zone	Binary	Yes: 2, no: 98	0.0
INJ_SEV	Injury severity (Dependent Variable)	Binary	Low: 79, high: 21	0.0

[1]For numeric variables: mean (st. dev.); for binary or nominal variables: % frequency of the top two classes.

TABLE 5.3 Tabulation of All Prediction Results Based on 10-fold Cross-Validation

Model Type		Confusion Matrices Low	Confusion Matrices High	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Artificial neural networks (ANN)	Low	12,864	1,464	85.77	81.31	89.78	0.865
	High	2,409	10,477				
Support vector machines (SVM)	Low	13,192	1,136	**90.41**	88.55	92.07	**0.928**
	High	1,475	11,411				
Decision trees (DT/C5)	Low	12,675	1,653	86.61	84.55	88.46	0.8790
	High	1,991	10,895				
Logistic regression (LR)	Low	8,961	2,742	76.97	77.27	76.57	0.827
	High	3,525	11,986				

and briefly describes the variables created and used for this study.

Table 5.3 shows the predictive accuracies of all four model types. It shows the confusion matrices, overall accuracy, sensitivity, specificity, and area under the receiver operating characteristics (ROC) curve measures obtained using 10-fold cross-validation for all four mode types. As the results indicate, SVM was the most accurate classification technique with better than 90 percent overall accuracy, comparably high sensitivity and specificity, and an area under the curve (AUC) value of 0.928 (of maximum 1.000). The next best model type was C5 decision tree algorithms with slightly better accuracy than ANN. The last in the accuracy ranking was LR, also with fairly good accuracy measures but not as good as the machine-learning methods.

Even though the accuracy measures obtained from all four model types were high enough to validate the proposed methodology, the main goal of this study was to identify and prioritize the significant risk factors influencing the level of injury severity sustained by drivers during a vehicle crash. To achieve this goal, a sensitivity analysis on all of the developed prediction models was conducted. Focusing on each model type individually, the variable importance measures for each fold were calculated using leave-one-out method, and then the results obtained were summed for each model type. To properly fuse (i.e., ensemble) the sensitivity analysis results for all four model types, the models' contribution to the fused/combined variable importance values were determined based on their cross-validation accuracy. That is, the best performing model type had the largest weight/contribution while the worst performing model type had the smallest weight/contribution. The fused variable importance values were tabulated, normalized, and then graphically presented in Figure 5.8.

Examination of the sensitivity analysis results revealed four somewhat distinct risk groups, each comprising four to eight variables. The top group, in an order from most to least importance, included REST_USE_N (whether the seat belt of any other restraining system was used), MANCOL_IM (manner of collision), EJECT_IM (whether the driver was ejected from the car), and DRUGRES1 (results of the drug test). According to the combined sensitivity analysis results of all prediction models, these four risk factors seemed to be significantly more important than the rest.

Questions for Case 5.3

1. What are the most important motivations behind analytically investigating car crashes?
2. How were the data in the Application Case acquired, merged, and reprocessed?

(*Continued*)

Application Case 5.3 (Continued)

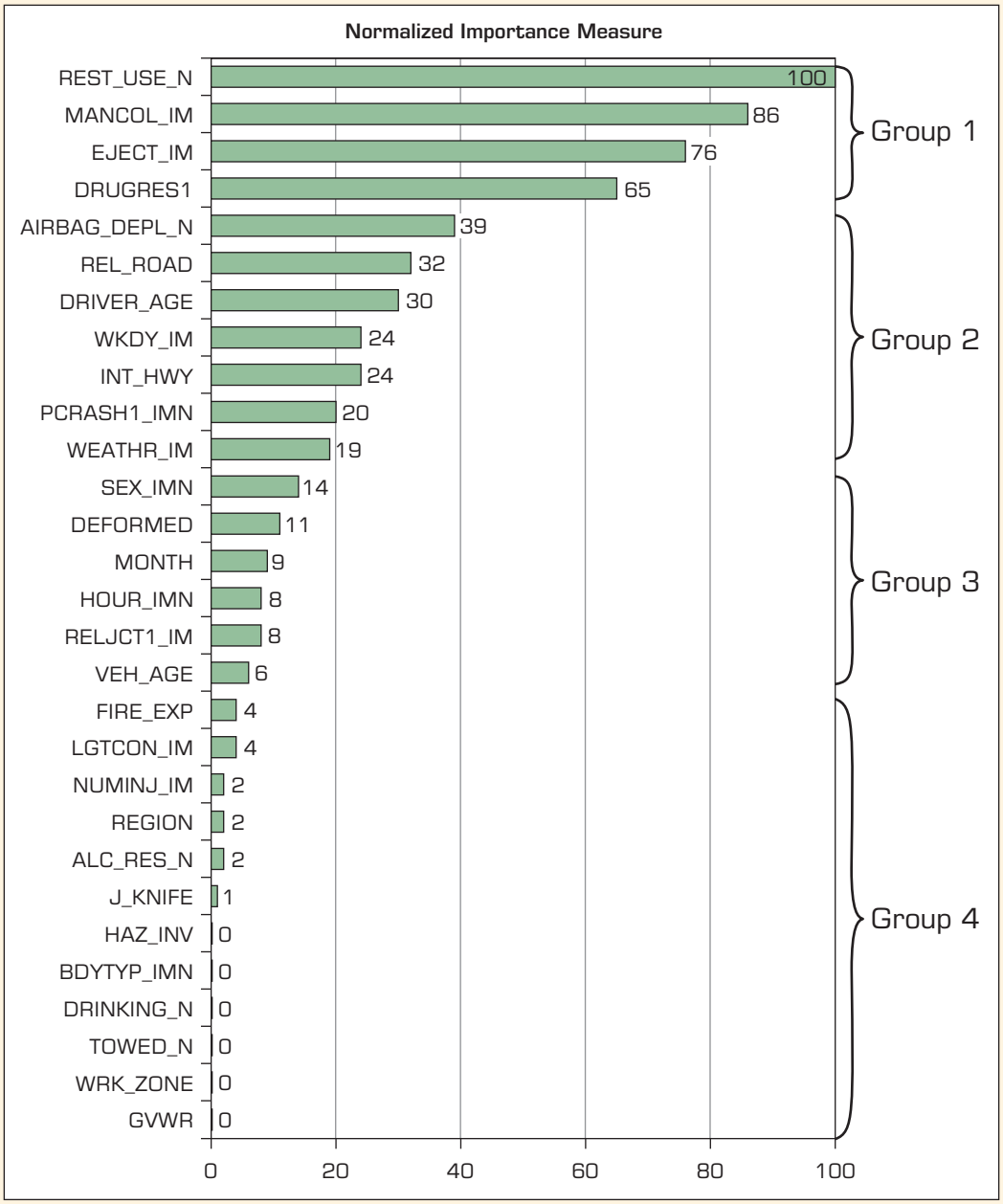


FIGURE 5.8 **Variable Importance Values.**

3. What were the results of this study? How can these findings be used for practical purposes?

Sources: D. Delen, L. Tomak, K. Topuz, & E. Eryarsoy, "Investigating Injury Severity Risk Factors in Automobile Crashes with Predictive Analytics and Sensitivity Analysis Methods," *Journal of Transport & Health, 4*, 2017, pp. 118–131; D. Friedman, "Oral Testimony Before the House Committee on Energy and Commerce, by the Subcommittee on Oversight and Investigations," April 1, 2014, **www.nhtsa.gov/Testimony** (accessed October 2017); National Highway Traffic Safety Administration (NHTSA's) (2018) General Estimate System (GES), **www.nhtsa.gov** (accessed January 20, 2018).

Mathematical Formulation of SVM

Consider data points in the training data set of the form:

$$\{(x_1,c_1), (x_2,c_2), \ldots, (x_n,c_n)\}$$

where the c is the class label taking a value of either 1 (i.e., "yes") or 0 (i.e., "no") while x is the input variable vector. That is, each data point is an m-dimensional real vector, usually of scaled $[0, 1]$ or $[-1, 1]$ values. The normalization and/or scaling are important steps to guard against variables/attributes with larger variance that might otherwise dominate the classification formulae. We can view this as training data, which denote the correct classification (something that we would like the SVM to eventually achieve) by means of a dividing hyperplane, which takes the mathematical form

$$w \cdot x - b = 0.$$

The vector w points perpendicularly to the separating hyperplane. Adding the offset parameter b allows us to increase the margin. In its absence, the hyperplane is forced to pass through the origin, restricting the solution. Because we are interested in the maximum margin, we also are interested in the support vectors and the parallel hyperplanes (to the optimal hyperplane) closest to these support vectors in either class. It can be shown that these parallel hyperplanes can be described by equations

$$w \cdot x - b = 1,$$
$$w \cdot x - b = -1.$$

If the training data are linearly separable, we can select these hyperplanes so that there are no points between them and then try to maximize their distance (see Figure 5.6b). By using geometry, we find the distance between the hyperplanes is $2/|w|$, so we want to minimize $|w|$. To exclude data points, we need to ensure that for all i either

$$w \cdot x_i - b \geq 1$$

or

$$w \cdot x_i - b \leq -1.$$

This can be rewritten as:

$$c_i(w \cdot x_i - b) \geq 1, \qquad 1 \leq i \leq n$$

Primal Form

The problem now is to minimize $|w|$ subject to the constraint $c_i(w \cdot x_i - b) \geq 1$, $1 \leq i \leq n$. This is a quadratic programming (QP) optimization problem. More clearly,

$$\text{Minimize } (1/2)\|w\|^2$$
$$\text{Subject to } c_i(w \cdot x_i - b) \geq 1, \quad 1 \leq i \leq n$$

The factor of $1/2$ is used for mathematical convenience.

Dual Form

Writing the classification rule in its dual form reveals that classification is only a function of the support vectors, that is, the training data that lie on the margin. The dual of the SVM can be shown to be:

$$\max \sum_{i=1}^{n} \alpha_i - \sum_{i,j} \alpha_i \alpha_j c_i c_j x_i^T x_j$$

where the α terms constitute a dual representation for the weight vector in terms of the training set:

$$w = \sum_i \alpha_i c_i x_i$$

Soft Margin

In 1995, Cortes and Vapnik suggested a modified maximum margin idea that allows for mislabeled examples. If there exists no hyperplane that can split the "yes" and "no" examples, the soft margin method will choose a hyperplane that splits the examples as cleanly as possible while still maximizing the distance to the nearest cleanly split examples. This work popularized the expression support vector machine or SVM. The method introduces slack variables, ξ_i, which measure the degree of misclassification of the datum.

$$c_i(w \cdot x_i - b) \geq 1 - \xi_i \quad 1 \leq i \leq n$$

The objective function is then increased by a function that penalizes non-zero ξ_i, and the optimization becomes a trade-off between a large margin and a small error penalty. If the penalty function is linear, the equation then transforms to

$$min \; \|w\|^2 + C\sum_i \xi_i \text{ such that } c_i(w \cdot x_i - \text{b}) \geq 1 - \xi_i \, 1 \leq i \leq n$$

This constraint along with the objective of minimizing $|w|$ can be solved using Lagrange multipliers. The key advantage of a linear penalty function is that the slack variables vanish from the dual problem with the constant C appearing only as a v-additional constraint on the Lagrange multipliers. Nonlinear penalty functions have been used, particularly to reduce the effect of outliers on the classifier, but unless care is taken, the problem becomes nonconvex, and thus it is considerably more difficult to find a global solution.

Nonlinear Classification

The original optimal hyperplane algorithm proposed by Vladimir Vapnik in 1963 while he was a doctoral student at the Institute of Control Science in Moscow was a linear classifier. However, in 1992, Boser, Guyon, and Vapnik suggested a way to create nonlinear classifiers by applying the kernel trick (originally proposed by Aizerman et al., 1964) to maximum-margin hyperplanes. The resulting algorithm is formally similar, except that every dot product is replaced by a nonlinear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in the transformed feature space. The transformation can be nonlinear and the transformed space high dimensional; thus, although the classifier is a hyperplane in the high-dimensional feature space, it can be nonlinear in the original input space.

If the kernel used is a Gaussian radial basis function, the corresponding feature space is a Hilbert space of infinite dimension. Maximum margin classifiers are well regularized, so the infinite dimension does not spoil the results. Some common kernels include:

Polynomial (homogeneous): $k(x, x') = (x \cdot x')$
Polynomial (inhomogeneous): $k(x, x') = (x \cdot x' + 1)$
Radial basis function: $k(x, x') = \exp(-\gamma\|x - x'\|^2)$, for $\gamma > 0$
Gaussian radial basis function: $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$

Sigmoid: $k(x, x') = \tan \text{h}\,(kx \cdot x' + c)$ for some $k > 0$ and $c < 0$

Kernel Trick

In machine learning, the kernel trick is a method for converting a linear classifier algorithm into a nonlinear one by using a nonlinear function to map the original observations into a higher-dimensional space; this makes a linear classification in the new space equivalent to nonlinear classification in the original space.

This is done using Mercer's theorem, which states that any continuous, symmetric, positive semi-definite kernel function $K(x, y)$ can be expressed as a dot product in a high-dimensional space. More specifically, if the arguments to the kernel are in a measurable space X and if the kernel is positive semi-definite—that is,

$$\sum_{i,j} K(x_i, x_j) c_i c_j \geq 0$$

for any finite subset $\{x_1, \ldots, x_n\}$ of X and subset $\{c_1, \ldots, c_n\}$ of objects (typically real numbers or even molecules)—then there exists a function $\varphi(x)$ whose range is in an inner product space of possibly high dimension, such that

$$K(x, y) = \varphi(x) \cdot \varphi(y)$$

The kernel trick transforms any algorithm that solely depends on the dot product between two vectors. Wherever a dot product is used, it is replaced with the kernel function. Thus, a linear algorithm can easily be transformed into a nonlinear algorithm. This nonlinear algorithm is equivalent to the linear algorithm operating in the range space of φ. However, because kernels are used, the φ function is never explicitly computed. This is desirable because the high-dimensional space could be infinite-dimensional (as is the case when the kernel is a Gaussian).

Although the origin of the term *kernel trick* is not known, it was first published by Aizerman et al. (Aizerman et al., 1964). It has been applied to several kinds of algorithm in machine learning and statistics, including:

- Perceptrons
- Support vector machines
- Principal components analysis
- Fisher's linear discriminant analysis
- Clustering

SECTION 5.4 REVIEW QUESTIONS

1. How do SVM work?
2. What are the advantages and disadvantages of SVM?
3. What is the meaning of "maximum-margin hyperplanes"? Why are they important in SVM?
4. What is the "kernel trick"? How is it used in SVM?

5.5 PROCESS-BASED APPROACH TO THE USE OF SVM

Due largely to the better classification results, SVM recently have become a popular technique for classification-type problems. Even though people consider them as being easier to use than artificial neural networks, users who are not familiar with the intricacies of SVM often get unsatisfactory results. In this section, we provide a process-based approach to the use of SVM, which is more likely to produce better results. A pictorial representation of the three-step process is given in Figure 5.9.

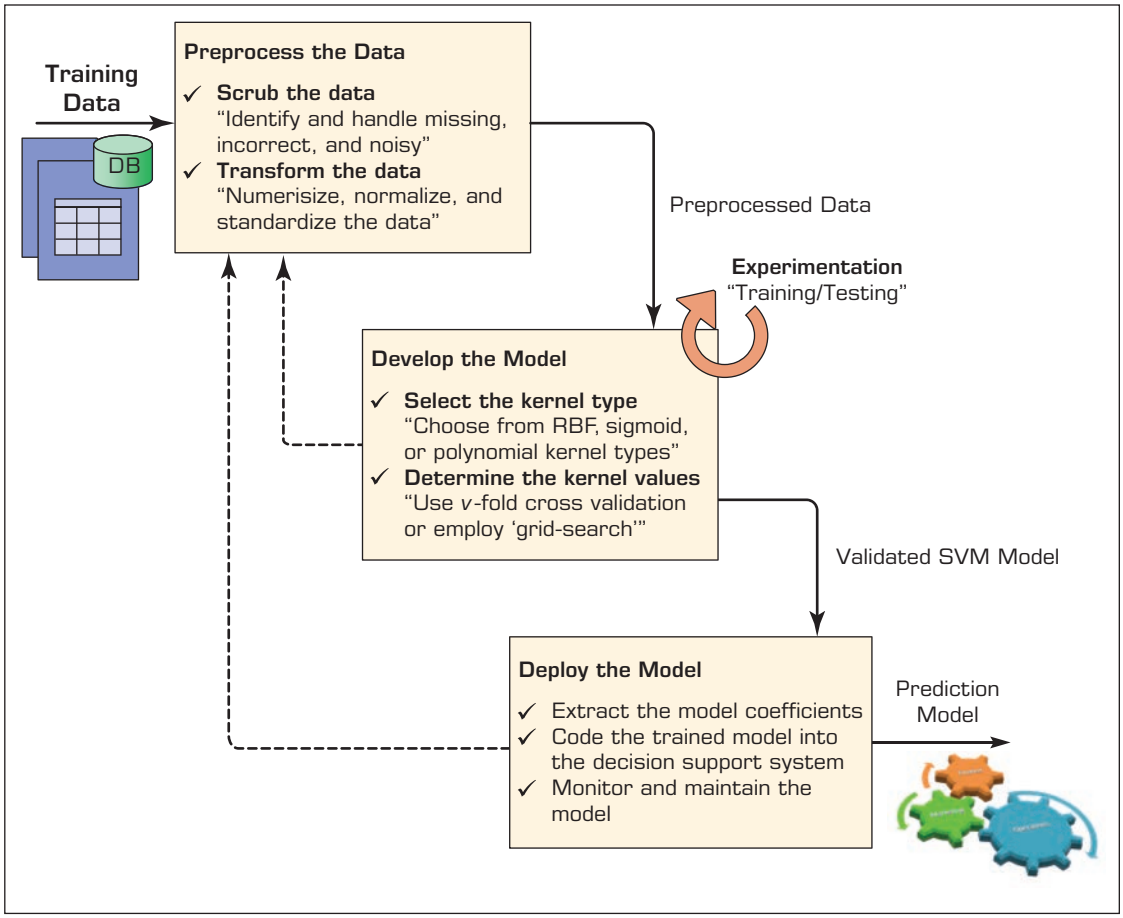


FIGURE 5.9 **Simple Process Description for Developing SVM Models.**

NUMERICIZING THE DATA SVM require that each data instance be represented as a vector of real numbers. Hence, if there are categorical attributes, we first have to convert them into numeric data. A common recommendation is to use m pseudo-binary variables to represent an m-class attribute (where $m \geq 3$). In practice, only one of the m variables assumes the value of 1 and others assume the value of 0 based on the actual class of the case (this is also called 1-of-m representation). For example, a three-category attribute such as {red, green, blue} can be represented as (0,0,1), (0,1,0), and (1,0,0).

NORMALIZING THE DATA As was the case for artificial neural networks, SVM also require normalization and/or scaling of numerical values. The main advantage of normalization is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is that it helps performing numerical calculations during the iterative process of model building. Because kernel values usually depend on the inner products of feature vectors (e.g., the linear kernel and the polynomial kernel), large attribute values might slow the training process. Use recommendations to normalize each attribute to the range $[-1, +1]$ or $[0, 1]$. Of course, we have to use the same normalization method to scale testing data before testing.

SELECT THE KERNEL TYPE AND KERNEL PARAMETERS Even though there are only four common kernels mentioned in the previous section, one must decide which one to use

(or whether to try them all, one at a time, using a simple experimental design approach). Once the kernel type is selected, then one needs to select the value of penalty parameter C and kernel parameters. Generally speaking, Radial Basis Function (RBF) is a reasonable first choice for the kernel type. The RBF kernel aims to nonlinearly map data into a higher dimensional space; by doing so (unlike with a linear kernel), it handles the cases in which the relation between input and output vectors is highly nonlinear. Besides, one should note that the linear kernel is just a special case of RBF kernel. There are two parameters to choose for RBF kernels: C and γ. It is not known beforehand which C and γ are the best for a given prediction problem; therefore, some kind of parameter search method needs to be used. The goal for the search is to identify optimal values for C and γ so that the classifier can accurately predict unknown data (i.e., testing data). The two most commonly used search methods are cross-validation and grid search.

DEPLOY THE MODEL Once an "optimal" SVM prediction model has been developed, the next step is to integrate it into the decision support system. For that, there are two options: (1) converting the model into a computational object (e.g., a Web service, Java Bean, or COM object) that takes the input parameter values and provides output prediction and (2) extracting the model coefficients and integrating them directly into the decision support system. The SVM models are useful (i.e., accurate, actionable) only if the behavior of the underlying domain stays the same. For some reason, if it changes, so does the accuracy of the model. Therefore, one should continuously assess the performance of the models and decide when they no longer are accurate, and, hence, need to be retrained.

Support Vector Machines versus Artificial Neural Networks

Even though some people characterize SVM as a special case of ANN, most recognize them as two competing machine-learning techniques with different qualities. Here are a few points that help SVM stand out against ANN. Historically, the development of ANN followed a heuristic path with applications and extensive experimentation preceding theory. In contrast, the development of SVM involved sound statistical learning theory first and then implementation and experiments. A significant advantage of SVM is that while ANN could suffer from multiple local minima, the solutions to SVM are global and unique. Two more advantages of SVM are that they have a simple geometric interpretation and give a sparse solution. The reason that SVM often outperform ANN in practice is that they successfully deal with the "over fitting" problem, which is a big issue with ANN.

Although SVM have these advantages (from a practical point of view), they have some limitations. An important issue that is not entirely solved is the selection of the kernel type and kernel function parameters. A second and perhaps more important limitation of SVM involves its speed and size, both in the training and testing cycles. Model building in SVM involves complex and time-demanding calculations. From the practical point of view, perhaps the most serious problem with SVM is the high algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks. Despite these limitations, because SVM are based on a sound theoretical foundation and the solutions they produce are global and unique in nature (as opposed to getting stuck in a suboptimal alternative such as a local minimum), today they are arguably among the most popular prediction modeling techniques in the data mining arena. Their use and popularity will only increase as the popular commercial data mining tools start to incorporate them into their modeling arsenal.

SECTION 5.5 REVIEW QUESTIONS

1. What are the main steps and decision points in developing an SVM model?
2. How do you determine the optimal kernel type and kernel parameters?

3. Compared to ANN, what are the advantages of SVM?
4. What are the common application areas for SVM? Search the Internet to identify popular application areas and specific SVM software tools used in those applications.

5.6 NEAREST NEIGHBOR METHOD FOR PREDICTION

Data mining algorithms tend to be highly mathematical and computationally intensive. The two popular ones that are covered in the previous section (i.e., ANN and SVM) involve time-demanding, computationally intensive iterative mathematical derivations. In contrast, the **k-nearest neighbor** algorithm (or *k*NN in short) seems overly simplistic for a competitive prediction method. What it does and how it does it are so easy to understand (and explain to others). *k*-NN is a prediction method for classification—as well as regression-type prediction problems. *k*-NN is a type of instance-based learning (or lazy learning) since the function is approximated only local and all computations are deferred until the actual prediction.

The *k*-nearest neighbor algorithm is among the simplest of all machine-learning algorithms: For instance, in the classification-type prediction, a case is classified by a majority vote of its neighbors with the object being assigned to the class most common among its k nearest neighbors (where k is a positive integer). If $k = 1$, then the case is simply assigned to the class of its nearest neighbor. To illustrate the concept with an example, let us look at Figure 5.10 where a simple two-dimensional space represents the values for the two variables (x, y); the star represents a new case (or object); and circles and squares represent known cases (or examples). The task is to assign the new case to either circles or squares based on its closeness (similarity) to one or the other. If you set the value of k to 1 ($k = 1$), the assignment should be made to square because the closest example to star is a square. If you set the value of k to 3 ($k = 3$), the assignment should be made to circle because there two circles and one square; hence, from the simple majority vote rule, the circle gets the assignment of the new case. Similarly, if you set the value of k to 5 ($k = 5$), then the assignment should be made to square class. This overly simplified example is meant to illustrate the importance of the value that one assigns to k.

The same method can be used for regression-type prediction tasks by simply averaging the values of its k nearest neighbors and assigning this result to the case being predicted. It can be useful to weight the contributions of the neighbors so that the nearer

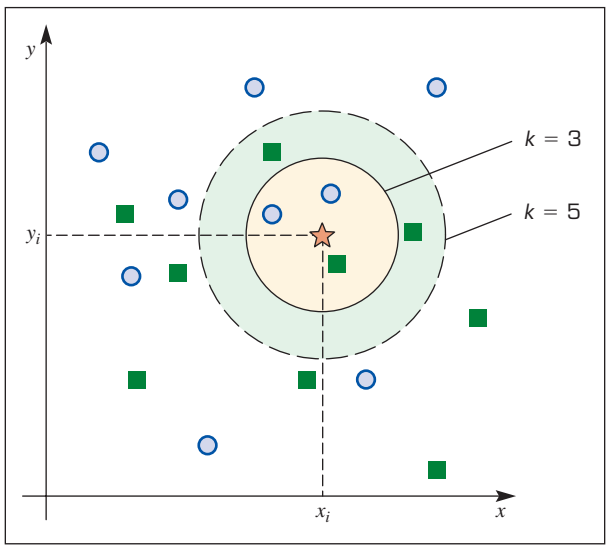


FIGURE 5.10 The Importance of the Value of k in kNN Algorithm.

neighbors contribute more to the average than the more distant ones. A common weighting scheme is to give each neighbor a weight of $1/d$, where d is the distance to the neighbor. This scheme is essentially a generalization of linear interpolation.

The neighbors are taken from a set of cases for which the correct classification (or, in the case of regression, the numerical value of the output value) is known. This can be thought of as the training set for the algorithm even though no explicit training step is required. The *k*-nearest neighbor algorithm is sensitive to the local structure of the data.

Similarity Measure: The Distance Metric

One of the two critical decisions that an analyst has to make while using *k*NN is to determine the similarity measure (the other is to determine the value of k, which is explained next). In the *k*NN algorithm, the similarity measure is a mathematically calculable distance metric. Given a new case, *k*NN makes predictions based on the outcome of the k neighbors closest in distance to that point. Therefore, to make predictions with *k*NN, we need to define a metric for measuring the distance between the new case and the cases from the examples. One of the most popular choices to measure this distance is known as Euclidean (Eq. 2), which is simply the linear distance between two points in a dimensional space; the other popular one is the rectilinear (a.k.a. city-block or Manhattan distance) (Eq. 3). Both of these distance measures are special cases of Minkowski distance (Eq. 1).

Minkowski distance

$$d(i, j) = \sqrt{\left(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q\right)} \qquad \text{(Eq. 1)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects (e.g., a new case and an example in the data set), and q is a positive integer.

If $q = 1$, then d is called Manhattan distance

$$d(i, j) = \sqrt{|x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|} \qquad \text{(Eq. 2)}$$

If $q = 2$, then d is called Euclidean distance

$$d(i, j) = \sqrt{\left(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2\right)} \qquad \text{(Eq. 3)}$$

Obviously, these measures apply only to numerically represented data. What about nominal data? There are ways to measure distance for non-numerical data as well. In the simplest case, for a multi-value nominal variable, if the value of that variable for the new case and that for the example case are the same, the distance would be 0, otherwise 1. In cases such as text classification, more sophisticated metrics exist, such as the overlap metric (or Hamming distance). Often, the classification accuracy of *k*NN can be improved significantly if the distance metric is determined through an experimental design in which different metrics are tried and tested to identify the best one for the given problem.

Parameter Selection

The best choice of k depends upon the data. Generally, larger values of k reduce the effect of noise on the classification (or regression) but also make boundaries between classes less distinct. An "optimal" value of k can be found by some heuristic techniques, for instance, cross-validation. The special case in which the class is predicted to be the class of the closest training sample (i.e., when $k = 1$) is called the *nearest neighbor algorithm*.

CROSS-VALIDATION Cross-validation is a well-established experimentation technique that can be used to determine *optimal* values for a set of unknown model parameters.

It applies to most, if not all, of the machine-learning techniques that have a number of model parameters to be determined. The general idea of this experimentation method is to divide the data sample into a number of randomly drawn, disjointed subsamples (i.e., v number of folds). For each potential value of k, the kNN model is used to make predictions on the v^{th} fold while using the $v - 1$ folds as the examples and to evaluate the error. The common choice for this error is the root-mean-squared-error (RMSE) for regression-type predictions and percentage of correctly classified instances (i.e., hit rate) for the classification-type predictions. This process of testing each fold against the remaining examples repeats v times. At the end of the v number of cycles, the computed errors are accumulated to yield a goodness measure of the model (i.e., how well the model predicts with the current value of the k). At the end, the k value that produces the smallest overall error is chosen as the optimal value for that problem. Figure 5.11 shows a simple process that uses the training data to determine *optimal* values for k and *distance metric*, which are then used to predict new incoming cases.

As we observed in the simple example given earlier, the accuracy of the kNN algorithm can be significantly different with different values of k. Furthermore, the predictive power of the kNN algorithm degrades with the presence of noisy, inaccurate, or irrelevant features. Much research effort has been put into feature selection and normalization/scaling to ensure reliable prediction results. A particularly popular approach is the use of evolutionary algorithms (e.g., genetic algorithms) to optimize the set of features included in the kNN prediction system. In binary (two-class) classification problems, it is helpful to choose k to be an odd number because this would avoid tied votes.

A drawback to the basic majority voting classification in kNN is that the classes with the more frequent examples tend to dominate the prediction of the new vector because they tend to come up in the k nearest neighbors when the neighbors are computed due to their large number. One way to overcome this problem is to weigh the classification taking into account the distance from the test point to each of its k nearest neighbors. Another way to overcome this drawback is by one level of abstraction in data representation.

The naïve version of the algorithm is easy to implement by computing the distances from the test sample to all stored vectors, but it is computationally intensive, especially when the size of the training set grows. Many nearest neighbor search algorithms have

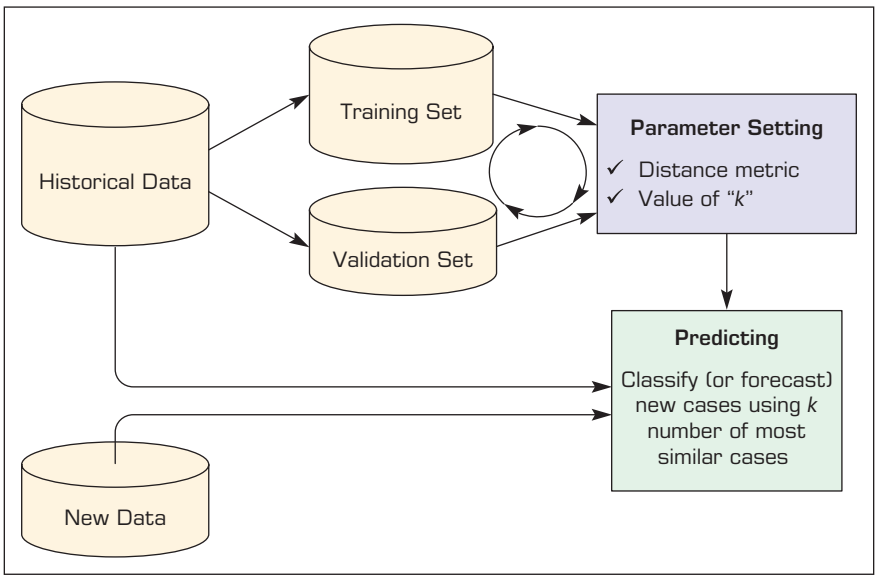


FIGURE 5.11 **Process of Determining the Optimal Values for Distance Metric and** *k*.

been proposed over the years; these generally seek to reduce the number of distance evaluations actually performed. Using an appropriate nearest neighbor search algorithm makes *k*NN computationally tractable even for large data sets. Refer to Application Case 5.4 about the superior capabilities of *k*NN in image recognition and categorization.

Application Case 5.4 Efficient Image Recognition and Categorization with *k*nn

Image recognition is an emerging data mining application field involved in processing, analyzing, and categorizing visual objects such as pictures. In the process of recognition (or categorization), images are first transformed to a multidimensional feature space and then, using machine-learning techniques, are categorized into a finite number of classes. Application areas of image recognition and categorization range from agriculture to homeland security, personalized marketing to environmental protection. Image recognition is an integral part of an artificial intelligence field called *computer vision*. As a technological discipline, computer vision seeks to develop computer systems that are capable of "seeing" and reacting to their environment. Examples of applications of computer vision include systems for process automation (industrial robots), navigation (autonomous vehicles), monitoring/detecting (visual surveillance), searching and sorting visuals (indexing databases of images and image sequences), engaging (computer–human interaction), and inspection (manufacturing processes).

While the field of visual recognition and category recognition has been progressing rapidly, much remains to be done to reach human-level performance. Current approaches are capable of dealing with only a limited number of categories (100 or so) and are computationally expensive. Many machine-learning techniques (including ANN, SVM, and *k*NN) are used to develop computer systems for visual recognition and categorization. Although commendable results have been obtained, generally speaking, none of these tools in their current form is capable of developing systems that can compete with humans.

Several researchers from the Computer Science Division of the Electrical Engineering and Computer Science Department at the University of California–Berkeley used an innovative ensemble approach to image categorization (Zhang et al., 2006). They considered visual category recognition in the framework of measuring similarities, or perceptual distances, to develop examples of categories. The recognition and categorization approach the researchers used was quite flexible, permitting recognition based on color, texture, and particularly shape. While nearest neighbor classifiers (i.e., *k*NN) are natural in this setting, they suffered from the problem of high variance (in bias-variance decomposition) in the case of limited sampling. Alternatively, one could choose to use SVM, but they also involve time-consuming optimization and computations. The researchers proposed a hybrid of these two methods, which deals naturally with the multiclass setting, has reasonable computational complexity both in training and at run time, and yields excellent results in practice. The basic idea was to find close neighbors to a query sample and train a local support vector machine that preserves the distance function on the collection of neighbors.

The researchers' method can be applied to large, multiclass data sets when it outperforms nearest neighbor and SVM and remains efficient when the problem becomes intractable. A wide variety of distance functions were used, and their experiments showed state-of-the-art performance on a number of benchmark data sets for shape and texture classification (MNIST, USPS, CUReT) and object recognition (Caltech-101).

Another group of researchers (Boiman and Irani, 2008) argued that two practices commonly used in image classification methods (namely, SVM- and ANN-type model-driven approaches and *k*NN-type nonparametric approaches) have led to less-than-desired performance outcomes. These researchers also claimed that a hybrid method can improve the performance of image recognition and categorization. They proposed a trivial Naïve Bayes

(*Continued*)

Application Case 5.4 (Continued)

*k*NN-based classifier, which employs *k*NN distances in the space of the local image descriptors (not in the space of images). The researchers claimed that, although the modified *k*NN method is extremely simple and efficient and requires no learning/training phase, its performance ranks among the top leading learning-based parametric image classifiers. Empirical comparisons of their method were shown on several challenging image categorization databases (Caltech-101, Caltech-256, and Graz-01).

In addition to image recognition and categorization, *k*NN is successfully applied to complex classification problems, such as content retrieval (handwriting detection, video content analysis, body and sign language (where communication is done using body or hand gestures), gene expression (another area in which *k*NN tends to perform better than other state-of-the-art techniques; in fact, a combination of *k*NN-SVM is one of the most popular techniques used here), and protein-to-protein interaction and 3D structure prediction (graph-based *k*NN is often used for interaction structure prediction).

Questions for Case 5.4

1. Why is image recognition/classification a worthy but difficult problem?
2. How can *k*NN be effectively used for image recognition/classification applications?

Sources: H. Zhang, A. C. Berg, M. Maire, & J. Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition," *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06),* Vol. 2, 2006, pp. 2126–2136; O. Boiman, E. Shechtman, & M. Irani, "In Defense of Nearest-Neighbor Based Image Classification," *IEEE Conference on Computer Vision and Pattern Recognition, 2008 (CVPR),* 2008, pp. 1–8.

SECTION 5.6 REVIEW QUESTIONS

1. What is special about the *k*NN algorithm?
2. What are the advantages and disadvantages of *k*NN as compared to ANN and SVM?
3. What are the critical success factors for a *k*NN implementation?
4. What is a similarity (or distance) measure? How can it be applied to both numerical and nominal valued variables?
5. What are the common applications of *k*NN?

5.7 NAÏVE BAYES METHOD FOR CLASSIFICATION

Naïve Bayes is a simple probability-based classification method (a machine-learning technique that is applied to classification-type prediction problems) derived from the well-known Bayes theorem. The method requires the output variable to have nominal values. Although the input variables can be a mix of numeric and nominal types, the numeric output variable needs to be discretized via some type of binning method before it can be used in a Bayes classifier. The word "Naïve" comes from its strong, somewhat unrealistic, assumption of independence among the input variables. Simply put, a Naïve Bayes classifier assumes that the input variables do not depend on each other, and the presence (or absence) of a particular variable in the mix of the predictors does not have anything to do with the presence or absence of any other variables.

Naïve Bayes classification models can be developed very efficiently (rather rapidly with very little computational effort) and effectively (quite accurately) in a supervised machine-learning environment. That is, by using a set of training data (not necessarily very large), the parameters for Naïve Bayes classification models can be obtained using the maximum likelihood method. In other words, because of the independence assumption, we can develop Naïve Bayes models without strictly complying with all of the rules and requirements of Bayes theorem. First let us review the Bayes theorem.

Bayes Theorem

To appreciate Naïve Bayes classification method, one would need to understand the basic definition of the Bayes theorem and the exact Bayes classifier (the one without the strong "Naïve" independence assumption). The Bayes theorem (also called *Bayes Rule*), named after the British mathematician Thomas Bayes (1701–1761), is a mathematical formula for determining conditional probabilities (the formula follows). In this formula, Y denotes the hypothesis and X denotes the data/evidence. This vastly popular theorem/rule provides a way to revise/improve prediction probabilities by using additional evidence.

The following formula shows the relationship between the probabilities of two events, Y and X. $P(Y)$ is the prior probability of Y. It is "prior" in the sense that it does not take into account any information about X. $P(X/Y)$ is the conditional probability of Y, given X. It is also called the *posterior probability* because it is derived from (or depends upon) the specified value of X. $P(X/Y)$ is the conditional probability of X given Y. It is also called the *likelihood*. $P(X)$ is the prior probability of X, which is also called *the evidence*, and acts as the normalizing constant.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad \rightarrow \quad Posterior = \frac{Likelihood * Prior}{Evidence}$$

$P(Y|X)$: *Posterior probability of Y given X*

$P(X|Y)$: *Conditional probability of X given Y (likelihood)*

$P(Y)$: *Prior probability of Y*

$P(X)$: *Prior probability of X (evidence, or unconditional probability of X)*

To numerically illustrate these formulas, let us look at a simple example. Based on the weather report, we know that there is a 40 percent chance of rain on Saturday. From the historical data, we also know that if it rains on Saturday, there is a 10 percent chance it will rain on Sunday; and if doesn't rain on Saturday, there is an 80 percent chance it will rain on Sunday. Let us say that "Raining on Sunday" is event Y, and "Raining on Monday" is event X. Based on the description we can write the following:

$P(Y)$ = Probability of raining on Saturday = 0.40
$P(X|Y)$ = Probability of raining on Sunday if it rained on Saturday = 0.10
$P(X)$ = Probability of raining on Monday = Sum of the probability of "Raining on Saturday and Raining on Sunday" and "Not Raining on Saturday and Raining on Sunday" = 0.40 * 0.10 + 0.60 * 0.80 = 0.52

Now if we were to calculate the probability for "It rained on Saturday?" given that it "Rained on Sunday," we would use Bayes theorem. It would allow us to calculate the probability of an earlier event given the result of a later event.

$$P(X|Y) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{0.10 * 0.40}{0.52} = 0.0769$$

Therefore, in this example, if it rained on Sunday, there's a 7.69 percent chance that it rained on Saturday.

Naïve Bayes Classifier

The Bayes classifier uses the Bayes theorem without the simplifying strong independence assumption. In a classification-type prediction problem, the Bayes classifier works as follows: Given a new sample to classify, it finds all other samples exactly like it (i.e., all predictor variables having the same values as the sample being classified); determines the class labels that they all belong to; and classifies the new sample into the most representative class. If none of the samples has the exact value match with the new class, then the classifier

TABLE 5.4 Sample Data Set for the Classification-Type Prediction Methods

	Input Variables (*X*)				Output Variable (*Y*)
Sample No.	**Outlook**	**Temperature**	**Humidity**	**Windy**	**Play Golf**
1	Sunny	Hot	High	No	No
2	Overcast	Hot	High	No	Yes
3	Rainy	Cool	Normal	No	Yes
4	Rainy	Cool	Normal	Yes	No
5	Overcast	Cool	Normal	Yes	No
6	Sunny	Hot	High	No	No
7	Sunny	Hot	High	No	Yes
8	Rainy	Mild	Normal	No	Yes
9	Sunny	Mild	Normal	Yes	Yes

will fail in assigning the new sample into a class label (because the classifier could not find any strong evidence to do so). Here is a very simple example. Using the Bayes classifier, we are to decide whether to play golf (Yes or No) for the following situation (Outlook is Sunny, Temperature is Hot, Humidity is High, and Windy is No). Table 5.4 presents historical samples that will be used to illustrate the specifics of our classification process.

Based on the historical data, three samples seem to match the situation (sample numbers 1, 6, and 7 as highlighted in Table 5.4). Of the three, two of the samples have the class label "No" and one has the label "Yes." Because the majority of the matched samples indicated "No," the new sample/situation is to be classified as "No."

Now let us consider a situation in which Outlook is Sunny, Temperature is Hot, Humidity is High, and Windy is Yes. Because there is no sample matching this value set, the Bayes classifier will not return a result. To find exact matches, there needs to be a very big data set. Even for the big data sets, as the number of predictor variables increases, the possibility of not finding an exact match increases significantly. When the data set and the number of predictor variables get larger, so does the time it takes to search for an exact match. All of these are the reasons why the Naïve Bayes classifier, a derivative of the Bayes classifier, is often used in predictive analytics and data mining practices. In the Naïve Bayes classifier, the exact match requirement is no longer needed. The Naïve Bayes classifier treats each predictor variable as an independent contributor to the prediction of the output variable and, hence, significantly increases its practicality and usefulness as a classification-type prediction tool.

Process of Developing a Naïve Bayes Classifier

Similar to other machine-learning methods, Naïve Bayes employs a two-phase model development and scoring/deployment process: (1) training in which the model/parameters are estimated and (2) testing in which the classification/prediction is performed on new cases. The process can be described as follows.

Training phase

Step 1. Obtain the data, clean the data, and organize them in a flat file format (i.e., columns as variables and rows as cases).

Step 2. Make sure that the variables are a nominal; if not (i.e., if any one of the variables is numeric/continuous), the numeric variables need to go through a data transformation (i.e., converting the numerical variable into nominal types by using discretization, such as binning).

Step 3. Calculate the prior probability of all class labels for the dependent variable.

Step 4. Calculate the likelihood for all predictor variables and their possible values with respect to the dependent variable. In the case of mixed variable types (categorical and continuous), each variable's likelihood (conditional probability) is estimated with the proper method that applies to the specific variable type. Likelihoods for nominal and numeric predictor variables are calculated as follows:

- For categorical variables, the likelihood (the conditional probability) is estimated as the simple fraction of the training samples for the variable value with respect to the dependent variable.
- For numerical variables, the likelihood is calculated by (1) calculating the mean and variance for each predictor variable for each dependent variable value (i.e., class) and then (2) calculating the likelihood using the following formula:

$$P(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

Quite often, the continuous/numerical independent/input variables are discretized (using an appropriate binning method), and then the categorical variable estimation method is used to calculate the conditional probabilities (likelihood parameters). If performed properly, this method tends to produce better predicting Naïve Bayes models.

Testing Phase

Using the two sets of parameters produced in steps 3 and 4 in the training phase, any new sample can be put into a class label using the following formula:

$$Posterior = \frac{Prior\ x\ Likelihood}{Evidence}$$

$$P(C \mid F_1, \ldots, F_n) = \frac{P(C)\ P(F_1, \ldots, F_n \mid C)}{P(F_1, \ldots, F_n)}$$

Because the denominator is constant (the same for all class labels), we can remove it from the formula, leading to the following simpler formula, which is essentially nothing but the joint probability.

$$classify\ (f_1, \ldots, f_n) = \underset{C}{\text{argmax}}\ p(C = c) \prod_{i=1}^{n} p(F_i = f_i \mid C = c)$$

This is a simple example to illustrate these calculations. In this example, we use the same data as shown in Table 5.4. The goal is to classify the following case: given that Outlook is Sunny, Temperature is Hot, Humidity is High, and Windy is No, what would be the class for the dependent variable (Play = Yes or No)?

From the data, we can observe that Prior (*Yes*) = 5/9 and Prior (*No*) = 4/9.

For the Outlook variable, the likelihoods are Likelihood (No/Sunny) = 2/3; Likelihood (No/Overcast); = ½; Likelihood (No/Rainy) = 1/3. The likelihood values of the other variables (Temperature, Humidity, and Wind) can be determined/calculated similarly. Again, the case we are trying to classify is Outlook is Sunny, Temperature is Hot, Humidity is High, and Windy is No. The results are shown in Table 5.5.

TABLE 5.5 Naïve Bayes Classification Calculations

			Ratio		Fraction (%/100)	
			Play: Yes	Play: No	Play = Yes	Play = No
Likelihood	Outlook =	Sunny	1/3	2/3	0.33	0.67
	Temperature =	Hot	2/4	2/4	0.50	0.50
	Humidity =	High	2/4	2/4	0.50	0.50
	Wind =	No	4/6	2/6	0.67	0.33
		Prior	5/9	4/9	0.56	0.44
			Product (multiply all)[1]		**0.031**	0.025
			Divide by the evidence[2]		**0.070**	0.056

[1]Does not include the denominator/evidence for the calculations; hence, it is a partial calculation.

[2]Include the denominator/evidence. Because the evidence is the same for all class labels (i.e., Yes and No), it does not make a difference in the classification result because both of the measures indicate class label as Yes.

Based on the results shown in Table 5.5, the answer would be Play Golf = Yes because it produces a larger value, 0.031 (compared to 0.025 for "No") as per the joint probabilities (the simplified calculation without the inclusion of the denominator). If we were to use the full posterior formula for the two class labels, which requires the inclusion of the denominator in the calculations, we observe 0.07 for "Yes" and 0.056 for "No." Because the denominator is common to all class labels, it will change the numerical output but not the class assignment.

Although Naïve Bayes is not very commonly used in predictive analytics projects today (because of its relatively poor prediction performance in a wide variety of application domains), one of its extensions, namely Bayesian network (see the next section), is gaining surprisingly rapid popularity among data scientists in the analytics world.

Application Case 5.5 provides an interesting example when many predictive analytics techniques are used to determine the changing condition of Crohn's disease patients in order to better manage this debilitating chronic disease. Along with Naïve Bayes, several statistical and machine-learning methods were developed, tested, and compared. The best performing model was then used to explain the rank-ordered importance (i.e., relative contribution) of all independent variables used in predicting the disease progress.

Application Case 5.5

Predicting Disease Progress in Crohn's Disease Patients: A Comparison of Analytics Methods

Introduction and Motivation

Inflammatory bowel disease (IBD), which includes Crohn's disease and ulcerative colitis (UC), impacts 1.6 million Americans, according to the Crohn's and Colitis Foundation (**crohnscolitisfoundation.org**). Crohn's disease causes chronic inflammation and damages the gastrointestinal tract. It can impact any part of the gastrointestinal tract. The cause of the disease is not entirely known, but some knowledge from research suggests that it could be caused by a combination of factors that include genetic makeup, immune system, and environmental settings. Systems that can detect disease progression or early disease onset can help in optimal utilization of healthcare resources and can result in better patient outcomes. The goal of this case study was to use

electronic medical records (EMRs) to predict and explain inflammation in Crohn's disease patients.

The Methodology

The data used in this study were from one of the nation's largest EMR databases, Cerner Health Facts EMR. It houses rich and varied information related to patients, healthcare settings, costs, reimbursement type, and prescription ordered data from multiple healthcare providers and hospitals in the United States. Data stored in the EMR database consists of patient-level data that were captured when a patient visited hospitals, urgent care centers, specialty clinics, general clinics, and nursing homes. The Health Facts database contains patient-level de-identified longitudinal data that were time stamped. The database was organized in the data tables as shown in Table 5.6.

A high-level process flow of the research methodology is shown in Figure 5.12. Although the process flow diagram did not provide the details of each step, it gave a high-level view of the sequence of the steps performed in the current predictive modeling study using EMR data. The three model types shown in the diagram were selected based on their comparatively better performance over other machine-learning methods such as Naïve Bayes, nearest neighbor, and neural networks. Detailed steps of data balancing and data standardization are explained in the paper by Reddy, Delen, and Agrawal (2018).

The Results

Prediction results were generated using the test set applying the repeated 10 times run on the 10-fold cross-validation method. The performance of each model was assessed by the metric AUC, which was the preferred performance metric over the prediction accuracy because the ROC curve, which generated the AUC, compared the classifier performance across the entire range of class distributions and error costs and, hence, is widely accepted as the performance measure for machine-learning applications. The mean AUC from the 10 run on the 10-fold cross-validation was generated (and shown in Table 5.7) for the three final model types—logistic regression, regularized regression, and gradient boosting machines (GBM).

Upon generation of the AUC for 100 models, researchers performed a post hoc analysis of variance (ANOVA) test and applied Tukey's Honest Significant Difference (HSD) test for multiple comparison tests to determine which classifier method's performance differed from the others based on the AUC. The test results showed that the mean AUC for regularized regression and the logistic regression did not differ significantly. However, the AUC from regularized regression and logistic regression were significantly different from the GBM model as seen in Table 5.8.

The relative importance of the independent variables was computed by adding the total amount of decrease in Gini index by the splits over a given predictor, averaged across all trees specified in the GBM tuning parameter, 1,000 trees in this research. This average decrease in GINI was normalized to a 0−100 scale on which a higher number indicates a stronger predictor. The variable importance results are shown in Figure 5.13.

Relative importance was computed by adding the total amount of decrease in Gini index by the splits over a given predictor averaged across all trees specified in the GBM tuning parameter, 1,000 trees in this research. This average decrease in GINI was normalized to a 0−100. scale on which a higher number

TABLE 5.6 Metadata of the Tables Extracted from EMR Database

Data Set (table)	Description
Encounter	Encounters including demographics, billing, healthcare setting, payer type, etc.
Medication	Medication orders sent by the healthcare provider
Laboratory	Laboratory data including blood chemistry, hematology, and urinalysis
Clinical Event	Clinical events data containing information about various metrics including body mass index, smoking status, pain score, etc.
Procedure	Clinical procedures performed on the patient

(*Continued*)

Application Case 5.4 (Continued)

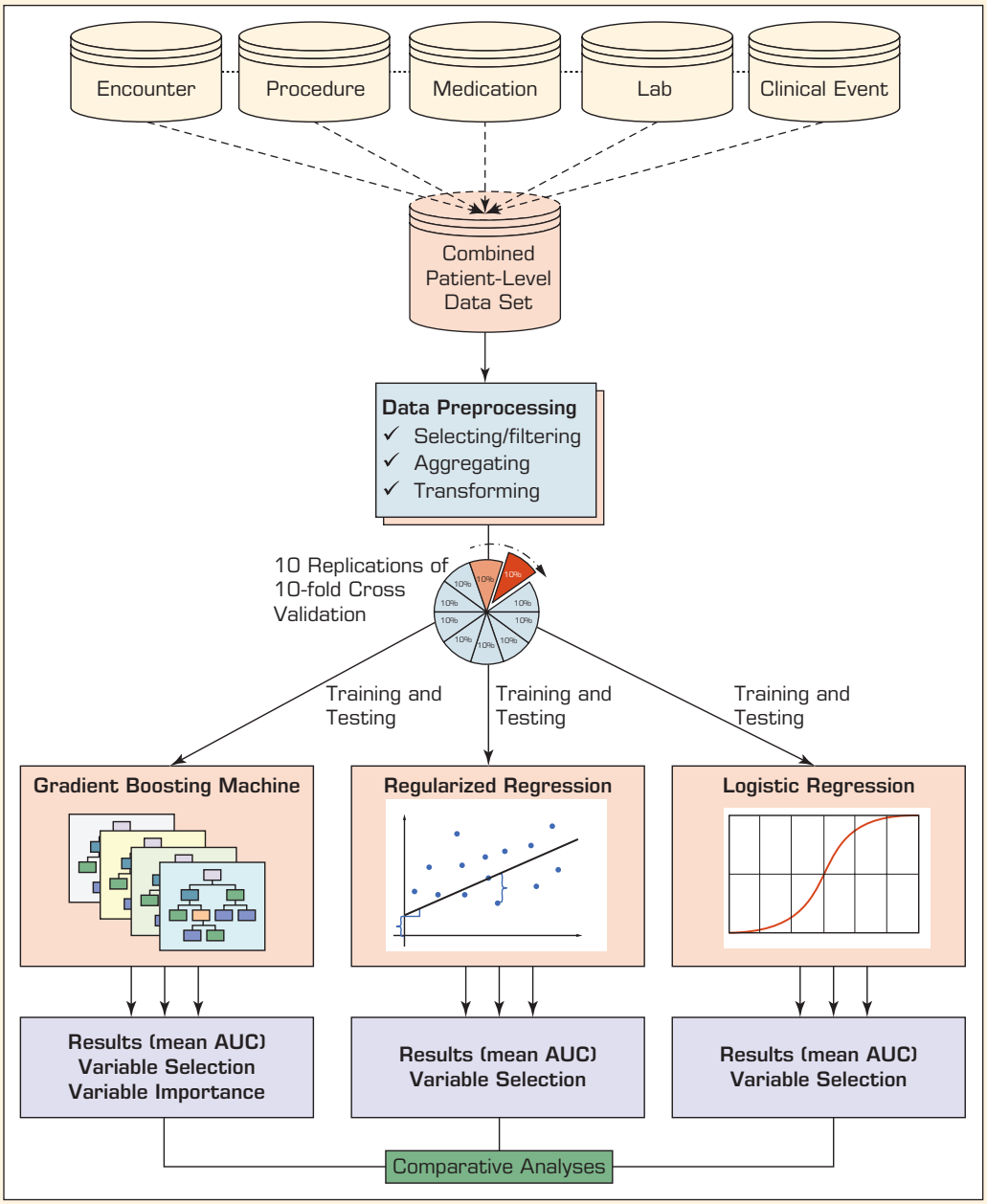


FIGURE 5.12 Process Flow Diagram of the High-Level Steps Involved in the Data Mining Research.

indicates a stronger predictor. The model results in Figure 5.13 showed that there was not one single predictor but a combination of predictors driving the predictions. Crohn's disease location at diagnosis such as small intestine and large intestine; lab parameters at baseline such as white blood cell (WBC) count; mean corpuscular hemoglobin (MCH); mean corpuscular volume; sodium; red blood cell (RBC); distribution of platelet count; creatinine; hematocrit; and hemoglobin were the strongest predictors. One of

TABLE 5.7 AUC for Each Repeated Run Across Three Models

Repeated Run	Logistic Regression	Regularized Regression	Gradient Boosting Machines (GBM)
1	0.7929	0.8267	0.9393
2	0.7878	0.8078	0.9262
3	0.8080	0.8145	0.9369
4	0.8461	0.8487	0.9124
5	0.8243	0.8281	0.9414
6	0.7681	0.8543	0.8878
7	0.8167	0.8154	0.9356
8	0.8174	0.8176	0.9330
9	0.8452	0.8281	0.9467
10	0.8050	0.8294	0.9230
Mean AUC	**0.8131**	**0.8271**	**0.9282**
Median AUC	**0.8167**	**0.8274**	**0.9343**

the strongest demographic predictors of the inflammation severity doubling was age. Other healthcare-setting and encounter-related variables such as hospital bed size, diagnosis priority, and region, whether south or not, predicting whether inflammation severity doubled or not, also had some predictive ability. The majority of the Crohn's disease researchers identified the location of the disease, age at diagnosis, smoking status, biologic markers, and tumor necrosis factor (TNF) levels to predict the response to treatment; these are some of the identifiers that also predicted the inflammation severity. Logistic regression and regularized regression cannot produce a similar relative variable importance plot. However, the odds ratio and standardized coefficients generated were used to identify the stronger predictors of inflammation severity.

This study was able to show that disease can be managed in real time by using decision support tools that rely on advanced analytics to predict the future inflammation state, which would then allow for medical intervention prospectively. With this information, healthcare providers can improve patient outcomes by intervening early and making

TABLE 5.8 ANOVA with Multiple Comparisons Using Tukey's Test

Tukey Grouping	Mean AUC	No. of Observations	Model Type
A	0.928	100	GBM
B	0.827	100	Regularized regression
B	0.812	100	Logistic regression

Means with the Same Letter Are Not Significantly Different

(*Continued*)

Application Case 5.5 (Continued)

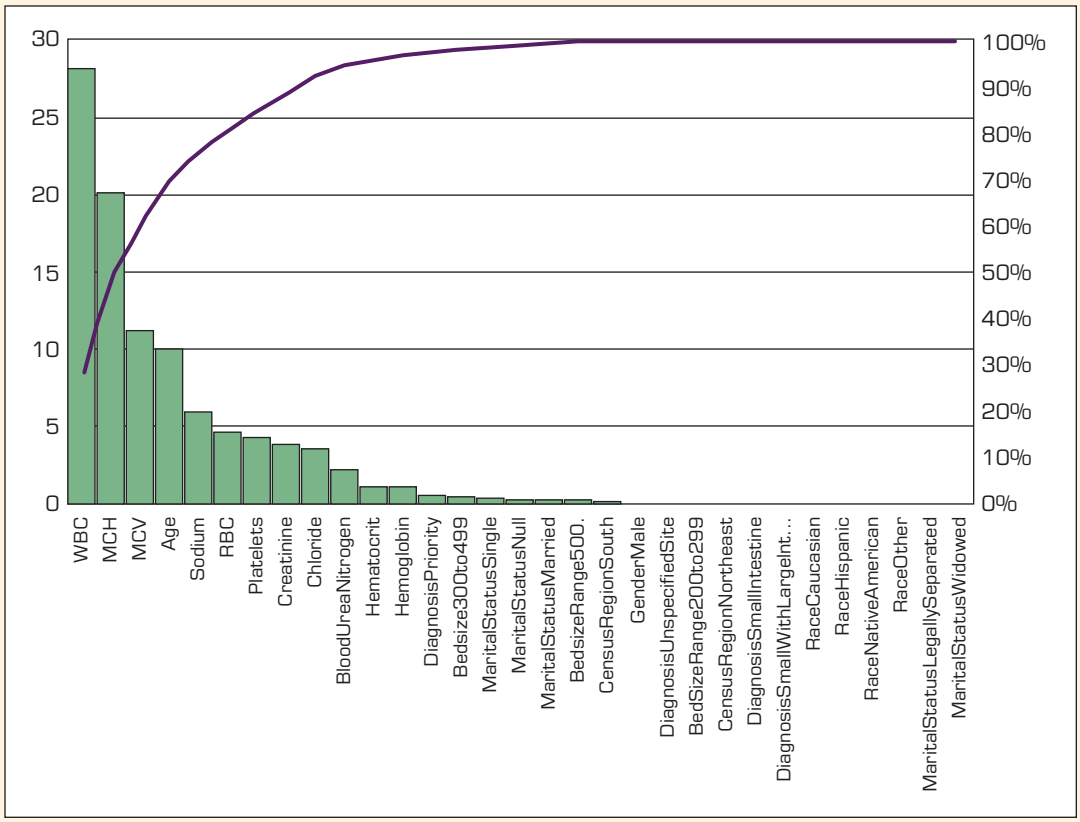


FIGURE 5.13 Relative Variable Importance for GBM Model.

necessary therapeutic adjustments that would work for the specific patient.

Questions for Case 5.5

1. What is Crohn's disease and why is it important?
2. Based on the findings of this Application Case, what can you tell about the use of analytics in chronic disease management?
3. What other methods and data sets might be used to better predict the outcomes of this chronic disease?

Source: B. K. Reddy, D. Delen, & R. K. Agrawal, "Predicting and Explaining Inflammation in Crohn's Disease Patients Using Predictive Analytics Methods and Electronic Medical Record Data," *Health Informatics Journal,* 2018.

SECTION 5.7 REVIEW QUESTIONS

1. What is special about the Naïve Bayes algorithm? What is the meaning of "Naïve" in this algorithm?
2. What are the advantages and disadvantages of Naïve Bayes compared to other machine-learning methods?
3. What type of data can be used in Naïve Bayes algorithm? What type of predictions can be obtained from it?
4. What is the process of developing and testing a Naïve Bayes classifier?

5.8 BAYESIAN NETWORKS

Bayesian belief networks or Bayesian networks (BN) were first defined in an early paper of Judea Pearl as "supportive of self-activated, multidirectional propagation of evidence that converges rapidly to a globally-consistent equilibrium" (Pearl, 1985). Later on, with his continuing work in this area, Pearl won the prestigious ACM's A.M. Turing Award for his contributions to the field of artificial intelligence and the development of BN. With this success, BN has received more public recognition than ever before, establishing it as a new paradigm in artificial intelligence, predictive analytics, and data science.

BN is a powerful tool for representing dependency structure in a graphical, explicit, and intuitive way. It reflects the various states of a multivariate model and their probabilistic relationships. Theoretically, any system can be modeled with BN. In a given model, some states will occur more frequently when others are also present; for example, if a freshman student is not registered for next fall (a presumed freshman student dropout case), the chances of the student's having financial aid is lower, indicating a relationship between the two variables. This is where the conditional probabilities (the basic theory that underlies BN) come to play to analyze and characterize the situation.

BNs have become popular among probabilistic graphical models because they have been shown to be able to capture and reason with complex, nonlinear, and partially uncertain situations and interactions (Koller and Friedman, 2009). While their solid, probability-based theoretical properties made Bayesian networks immediately attractive for academic research, especially for studying causality, their use in practical data science and business analytics domains is relatively new. For instance, researchers recently have developed data analytics–driven BN models in such domains that include predicting and understanding the graft survival for kidney transplantations (Topuz et al., 2018), predicting failures in the rail industry caused by weather-related issues (Wang et al., 2017), predicting food fraud type (Bouzembrak et al., 2016), and detecting diseases (Meyfroidt et al., 2009).

Essentially, the BN model is a directed acyclic graph whose nodes correspond to the variables and arcs that signify conditional dependencies between variables and their possible values (Pearl, 2009). Here is a simple example, which was previously used as Application Case 3.2. For details of the example, please reread this Application Case. Let us say that the goal was to predict whether a freshman student will stay or drop out of college (presented in the graph as SecondFallRegistered) using some data/information about the student such as (1) the declared college type (a number of states/options exists for potential colleges) and (2) whether the student received financial aid in the first fall semester (two states exists, Yes or No), both of which can be characterized probabilistically using the historical data. One might think that there exist some causal links among the three variables, both college type and financial aid relating to whether the student comes back for the second fall semester and that it is reasonable to think that some colleges historically have more financial support than others (see Figure 5.14 for the presumed causal relationships).

The direction of links in BN graphs corresponded to the probabilistic or conditional dependencies between any two variables. Calculating actual conditional probabilities using historical data would help predict and understand student retention (SecondFallRegistered) using two variables, "financial aid" and "college type." Such a network can then be used to answer questions such as these:

- Is the college type "engineering"?
- What are the chances the student will register next fall?
- How will financial aid affect the outcome?

How Does BN Work?

Building probabilistic models such as BN of complex real-world situations/problems using historical data can help in predicting what is likely to happen when something else would have happened. Essentially, BN typically tries to represent interrelationships

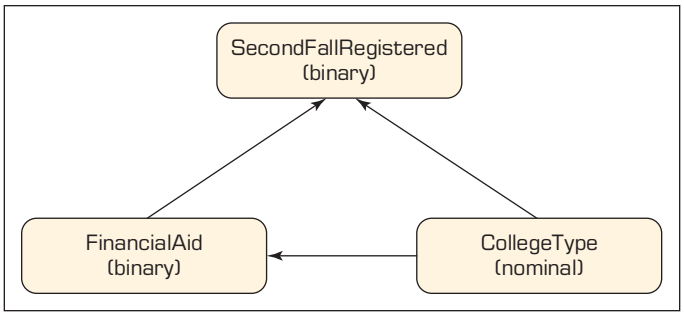


FIGURE 5.14 **Simple Illustration of the Partial Causality Relationships in Student Retention.**

among the variables (both input and output variables) using a probabilistic structure that is often called the *joint distribution*. Joint distributions can be presented as a table consisting of all possible combinations of states (variable values) in a given model. For complex models, such a table can easily become rather large because it stores one probability value for every combination of states. To mitigate the situation, BN does not connect all of nodes in the model to each other; rather, it connects only the nodes that are probabilistically related by some sort of conditional and/or logical dependency, resulting in significant savings on the computations.

The naturally complex probability distributions can be represented in a relatively compact way using BNs' conditional independence formula. In the following formula, each x_i represents a variable and $P_{a_{x_i}}$ represents the parents of that variable; by using these representations, the BN chain rule can be expressed as follows (Koller and Friedman, 2009):

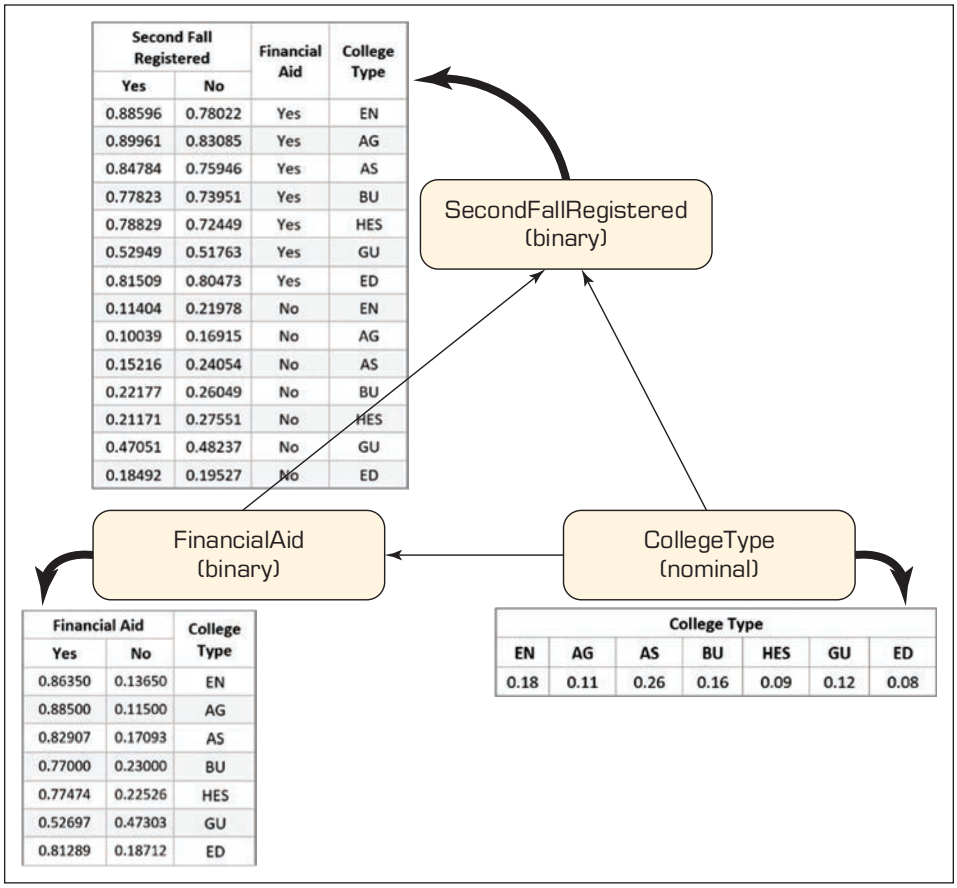
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Pa_{x_i}).$$

Let's look at an example by building a simple network for the student retention prediction problem. Remember that our problem is to predict whether a freshman student will stay for the second fall semester or drop out of college by using some data/information from student records (i.e., declared college type and whether the student received financial aid for the first semester). The constructed BN graphical model shown in Figure 5.15 exhibits the relationships and conditional probabilities among all three nodes.

How Can BN Be Constructed?

There are two common methods available to construct the network: (1) manually with the help of a domain expert and (2) analytically by learning the structure of the network from the historical data by using advanced mathematical methods. Building a network manually, even for a modest size network, requires a skilled knowledgeable engineer spending several hours with the domain expert. As the size of the network gets larger, the time spent by the engineer and the domain expert increases exponentially. In some cases, it is very difficult to find a knowledgeable expert for a particular domain. Even if such a domain expert exists, he or she might not have the time to devote to the model building effort and/or might not be explicit and articulate enough (i.e., explaining tacit knowledge is always a difficult task) to be of much use as a source of knowledge. Therefore, most of the previous studies developed and offered various techniques that can be used to learn the structure of the network automatically from the data.

One of the earlier methods used to learn the structure of the network automatically from the data is the Naïve Bayes method. The Naïve Bayes classification method is a simple probabilistic model that assumes conditional independence between all predictor



Second Fall Registered		Financial Aid	College Type
Yes	No		
0.88596	0.78022	Yes	EN
0.89961	0.83085	Yes	AG
0.84784	0.75946	Yes	AS
0.77823	0.73951	Yes	BU
0.78829	0.72449	Yes	HES
0.52949	0.51763	Yes	GU
0.81509	0.80473	Yes	ED
0.11404	0.21978	No	EN
0.10039	0.16915	No	AG
0.15216	0.24054	No	AS
0.22177	0.26049	No	BU
0.21171	0.27551	No	HES
0.47051	0.48237	No	GU
0.18492	0.19527	No	ED

Financial Aid		College Type
Yes	No	
0.86350	0.13650	EN
0.88500	0.11500	AG
0.82907	0.17093	AS
0.77000	0.23000	BU
0.77474	0.22526	HES
0.52697	0.47303	GU
0.81289	0.18712	ED

College Type						
EN	AG	AS	BU	HES	GU	ED
0.18	0.11	0.26	0.16	0.09	0.12	0.08

FIGURE 5.15 **Conditional Probability Tables for the Two Predictor and One Target Variables.**

variables and the given class/target variable to learn the structure. The classification algorithm is based on the Bayes rule that the probability of class/target value is computed for each given attribute variable and then the highest prediction is chosen for the structure.

A more recent and popular method for learning the structure of the network is called *Tree Augmented Naïve* (TAN) *Bayes*. The TAN method is an updated version of the Naïve Bayes classifier that uses tree structure to approximate the interactions between predictor variables and the target variable (Friedman, Geiger, and Goldszmidt, 1997). In the TAN model structure, class variable has no parent, and each and every predictor variable has the class variable as its parent along with at most one other predictor variable (i.e., attribute) as shown in Figure 5.16. Thus, an arc between two variables indicates a directional and causal relationship between them. Formal representation of parents for a variable x_i can be shown by:

$$Pa_{x_i} = \{C, x_{\delta(i)}\}$$

where the tree is a function over $\delta(i) > 0$, and $P_{a_{x_i}}$ is the set of parents for each x_i. A class variable (C) has no parents, namely $Pa_C = \varnothing$. It is empirically and theoretically shown that TAN performs better than Naïve Bayes and maintains simplicity in the computations because it does not require a search process (Friedman et al., 1997).

The procedure for constructing a TAN uses Chow and Liu's tree Bayesian concept. Finding a maximally weighted spanning tree in a graph is an optimization problem

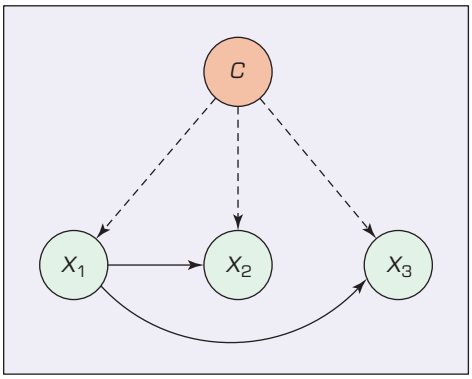


FIGURE 5.16 **Tree Augmented Naïve Bayes Network Structure.**

whose objective is to maximize log likelihood of $\delta(i)$ (Chow and Liu 1968). Then the TAN construction steps can be as described as follows (Friedman, et al., 1997):

Step 1. Compute the conditional mutual information function for each (i, j) pair as

$$I_P(x_i : x_j|C) = \sum_{x_i, x_j, C} P(x_i, x_j, C)\log\frac{P(x_i, x_j|C)}{P(x_i|C)P(x_j|C)},\ i \neq j$$

This function indicates how much information is provided when the class variable is known.

Step 2. Build a complete undirected graph and use a conditional mutual information function to annotate the weight of an edge connecting x_i to x_j.

Step 3. Build a maximum weighted spanning tree.

Step 4. Convert the undirected graph into a directed one by choosing a root variable and setting the direction of all edges to be outward from it.

Step 5. Construct a TAN model by adding a vertex labeled by C and an arc from C to each x_i.

One of the superior features of BN is its ease of adaptability. While building a BN, one can start the network as small with a limited knowledge and then expand on it as new information becomes available. In such situation, having missing values in the data set might not be a major issue because it can use the available portion of the data/values/knowledge to create the probabilities. Figure 5.17 shows a fully developed, data-driven BN example for the student retention project.

From the applicability perspective, such a fully constructed BN can be of great use to practitioners (i.e., administrators and managers in educational institutions) because it offers a holistic view of all relationships and provides the means to explore detailed information using a variety of "what-if" analysis. In fact, with this network model, it is possible to calculate the student-specific risk probability of attrition, which is the likelihood or posterior probability of the student who would drop out, by systematically selecting and changing the value of a predictor variable within its value domain (assessing how much the dropout risk of a student changes as the value of a given predictor variable such as Fall-GPA changes).

When interpreting the BN model shown in Figure 5.17, one should consider the arcs, directions of the arrows on those arcs, direct interactions, and indirect relationships. For example, the fall grant/tuition waiver/scholarship category

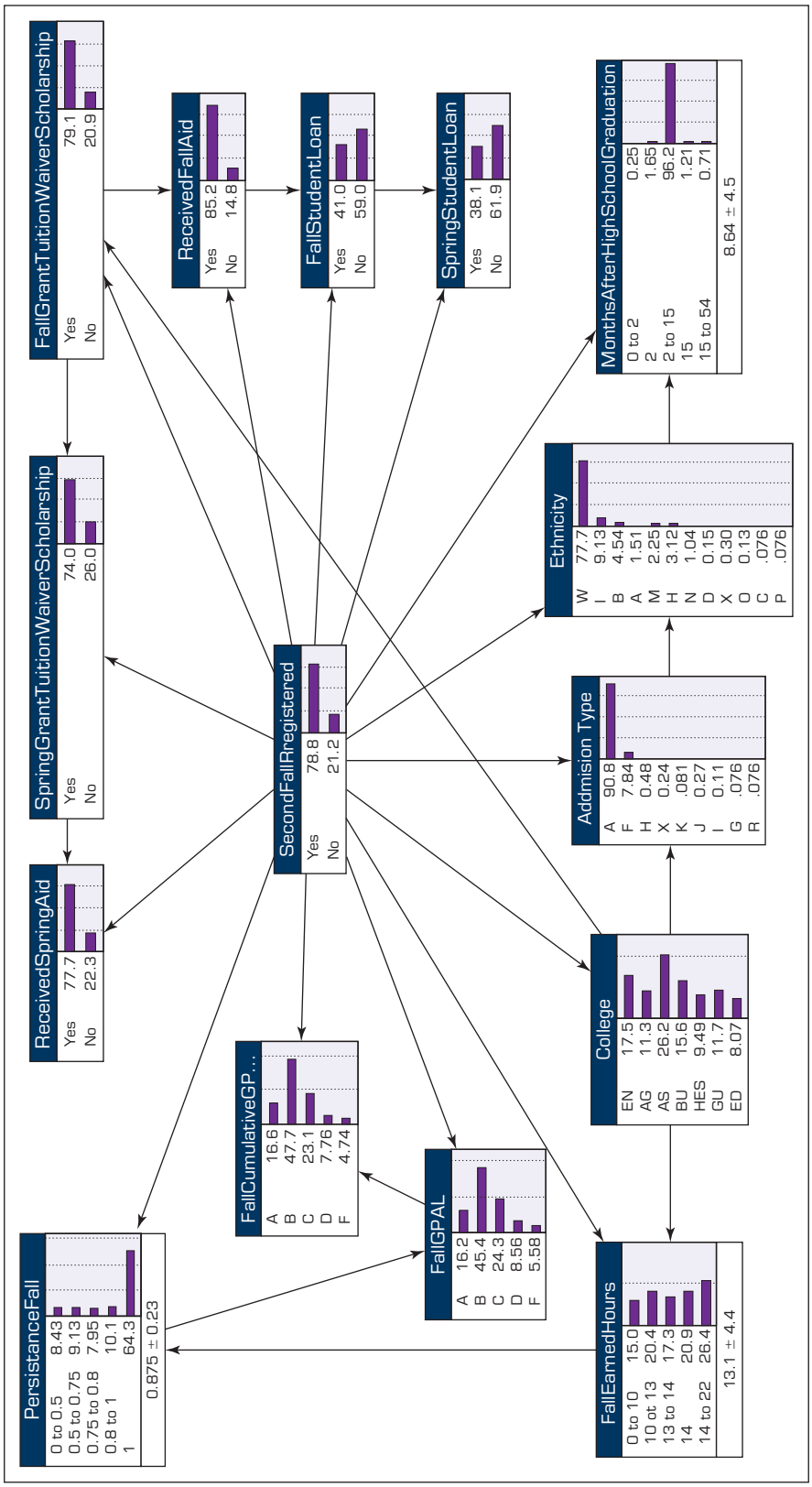


FIGURE 5.17 Bayesian Belief Network for Predicting Freshmen Student Attrition.

(i.e., FallGrantTuitionWaiverScholarship) and all the nodes linked to FallGrantTuition WaiverScholarship are related to student attrition (i.e., SecondFallRegistered). Moreover, while FallGrantTuitionWaiverScholarship interacts with college (College) and spring grant/tuition waiver/scholarship (i.e., SpringGrantTuitionWaiverScholarship) directly, it also interacts with admission type (AdmissionType) indirectly through College. According to the BN model, one of the most interactive predictors is the student's earned credit hours by registered rate (i.e., PersistanceFall), which contributes to the effect of the student's fall GPA (FallGPA) and student attrition. As such, if the PersistanceFall of the student is less than 0.8, then College type has an effect on student attrition. However, if the PersistanceFall of the student is 1.0, the College type does not impact the student attrition in a noteworthy manner.

As a collective view to what-if scenarios, Figure 5.18 summarizes the most positive and most negative levels within each predictor with its posterior probabilities. For instance, getting an A for the Fall GPA decreases the posterior probability of student attrition to 7.3 percent, or conversely, getting an F increases the probability of attrition to 87.8 percent where the baseline is 21.2 percent.

Some people have had doubts about using BN because they thought that BN does not work well if the probabilities upon which it is constructed are not exact. However, it turns out that in most cases, approximate probabilities driven from data and even the subjective ones that are guessed by domain experts provide reasonably good results. BN are shown to be quite robust toward imperfect, inaccurate, and incomplete knowledge. Often the combination of several strands of imperfect knowledge allows BN to make surprisingly good predictions and conclusions. Studies have shown that people are better at estimating probabilities "in the forward direction." For example, managers are quite good at providing probability estimates for "If the student has dropped out the college, what are the chances his or her college type is Art & Sciences?" rather than the reverse, "If the

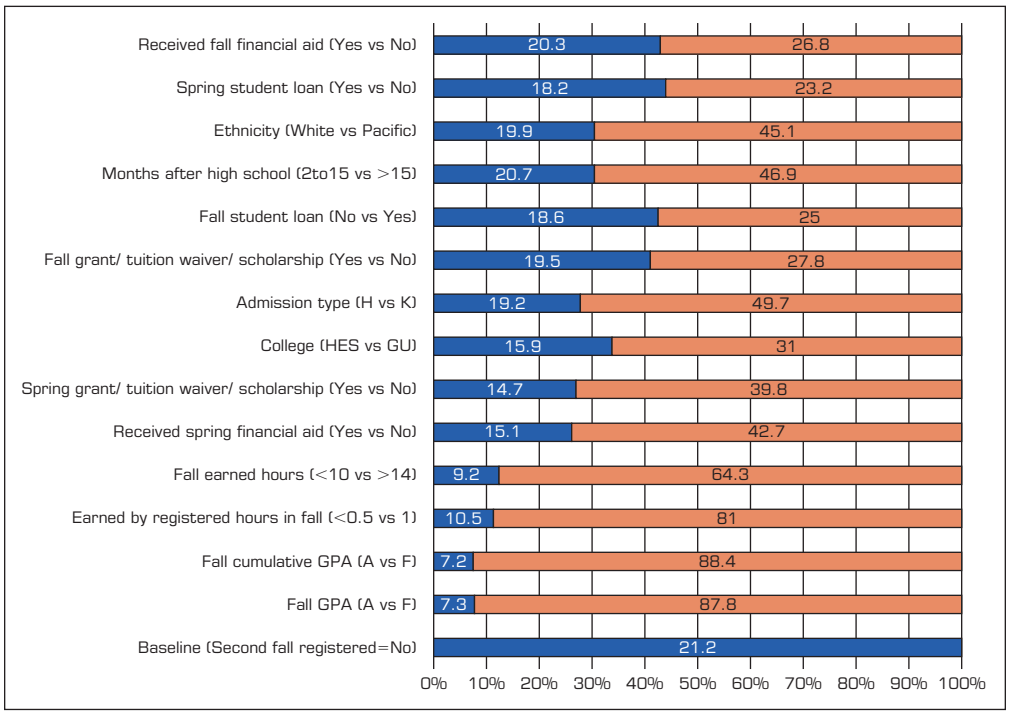


FIGURE 5.18 Probability of Student Attrition for Risk Factors—What-If Analysis on Individual Factors.

student goes to Art & Sciences college, what are the chances that this student will not register the next fall?"

SECTION 5.8 REVIEW QUESTIONS

1. What are Bayesian networks? What is special about them?
2. What is the relationship between Naïve Bayes and Bayesian networks?
3. What is the process of developing a Bayesian networks model?
4. What are the advantages and disadvantages of Bayesian networks compared to other machine-learning methods?
5. What is Tree Augmented Naïve (TAN) Bayes and how does it relate to Bayesian networks?

5.9 ENSEMBLE MODELING

Ensembles (or more appropriately called *model ensembles* or *ensemble modeling*) are combinations of the outcomes produced by two or more analytics models into a compound output. Ensembles are primarily used for prediction modeling when the scores of two or more models are combined to produce a better prediction. The prediction can be either classification or regression/estimation type (i.e., the former predicting a class label and the latter estimating a numerical output variable). Although use of ensembles has been dominated by prediction-type modeling, it can also be used for other analytics tasks such as clustering and association rule mining. That is, model ensembles can be used for supervised as well as unsupervised machine-learning tasks. Traditionally, these machine-learning procedures focused on identifying and building the best possible model (often the most accurate predictor of the holdout data) from a large number of alternative model types. To do so, analysts and scientists used an elaborate experimental process that mainly relied on trial and error to improve each single model's performance (defined by some predetermined metrics, e.g., prediction accuracy) to its best possible level so that the best of the models can be used/deployed for the task at hand. The ensemble approach turns this thinking around. Rather than building models and selecting the single best model to use/deploy, it proposes to build many models and use them all for the task they are intended to perform (e.g., prediction).

Motivation—Why Do We Need to Use Ensembles?

Usually researchers and practitioners build ensembles for two main reasons: for better accuracy and more stable/robust/consistent/reliable outcomes. Numerous research studies and publications over the past two decades have shown that ensembles almost always improve predictive accuracy for the given problem and rarely predict worse than the single models (Abbott, 2014). Ensembles began to appear in the data mining/analytics literature in 1990s, motivated by the limited success obtained by the earlier works on combining forecasts that dated a couple or more decades. By the early- to mid-2000s, ensembles had become popular and almost essential to winning data mining and predictive modeling competitions. One of the most popular awards for ensemble competitions is perhaps the famous Netflix prize, which was an open competition that solicited researchers and practitioners to predict user ratings of films based on historical ratings. The prize was US$1 million for a team that could reduce the RMSE of the then-existing Netflix internal prediction algorithm by the largest margin but no less than 10 percentage points. The winner, runner-up, and nearly all the teams at the top of the leaderboard used model ensembles in their submissions. As a result, the winning submission was the result of an ensemble containing hundreds of predictive models.

When it comes to justifying the use of ensembles, Vorhies (2016) put it the best—if you want to win a predictive analytics competition (at Kaggle or at anywhere else) or at least get a respectable place on the leaderboard, you need to embrace and intelligently use model ensembles. Kaggle has become the premier platform for data scientists to showcase their talents. According to Vorhies, the Kaggle competitions are like Formula One racing for data science. Winners edge out competitors at the fourth decimal place and, like Formula One race cars, not many of us would mistake them for daily drivers. The amount of time devoted and the extreme techniques used would not always be appropriate for an ordinary data science production project, but like paddle shifters and exotic suspensions, some of those improvements and advanced features find their way into the day-to-day life and practice of analytics professionals. In addition to Kaggle competitions, reputable organizations such as the Association for Computing Machinery (ACM)'s Special Interest Group (SIG) on Knowledge Discovery and Data Mining (SIGKDD) and Pacific-Asia Conference in Knowledge Discovery and Data Mining (PAKDD) regularly organize competitions (often called "cups") for the community of data scientists to demonstrate their competence, sometimes for monetary rewards but most often for simple bragging rights. Some of the popular analytics companies like the SAS Institute and Teradata Corporation organize similar competitions for (and extend a variety of relatively modest awards to) both graduate and undergraduate students in universities all over the world, usually in concert with their regular analytics conferences.

It is not just the accuracy that makes model ensembles popular and unavoidably necessary. It has been shown time and time again that ensembles can improve model accuracy, but they can also improve model robustness, stability, and, hence, reliability. This advantage of model ensembles is equally (or perhaps more) important and invaluable than accuracy in situations for which reliable prediction is of the essence. In ensemble models, by combining (some form of averaging) multiple models into a single prediction outcome, no single model dominates the final predicted value of the models, which in turn, reduces the likelihood of making a way-off-target "wacky" prediction. Figure 5.19 shows a graphical illustration of model ensembles for classification-type prediction problems. Although some varieties exist, most ensemble modeling methods

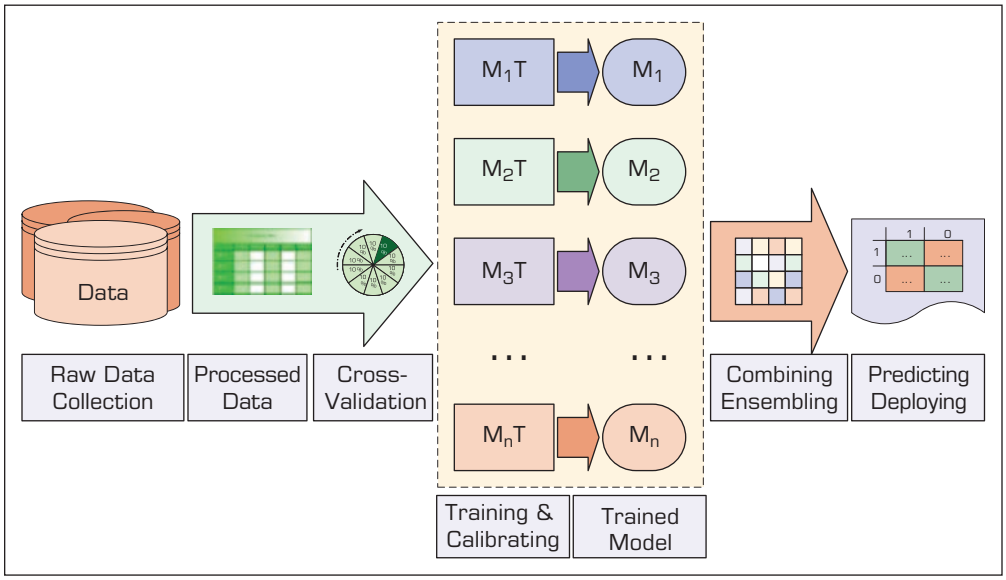


FIGURE 5.19 Graphical Depiction of Model Ensembles for Prediction Modeling.

follow this generalized process. From left to right, Figure 5.19 illustrates the general tasks of data acquisition and data preparation followed by cross-validation and model building and testing, and finally assembling/combining the individual model outcomes and assessing the resultant predictions.

Another way to look at ensembles is from the perspective of "collective wisdom" or "crowdsourcing." In the popular book *The Wisdom of Crowds* (Surowiecki, 2005), the author proposes that better decisions can be made if rather than relying on a single expert, many (even uninformed) opinions (obtained by a process called *crowdsourcing*) can be aggregated into a decision that is superior to the best expert's opinion. In his book, Surowiecki describes four characteristics necessary for the group opinion to work well and not degenerate into the opposite effect of poor decisions as evidenced by the "madness of crowds": diversity of opinion, independence, decentralization, and aggregation. The first three characteristics relate to how individual decisions are made—they must have information that differs from that of others in the group and is not affected by the others in the group. The last characteristic merely states that the decisions must be combined. These four principles/characteristics seem to lay the foundation for building better model ensembles as well. Each predictive model has a voice in the final decision. The diversity of opinion can be measured by the correlation of the predictive values themselves—if all of the predictions are highly correlated, or in other words, if the models nearly all agree, there is no foreseeable advantage in combining them. The decentralization characteristic can be achieved by resampling data or case weights; each model uses either different records from a common data set or at least uses the records with weights that differ from the other models (Abbott, 2014).

One of the prevalent concepts in statistics and predictive modeling that is highly relevant to model ensembles is the bias-variance trade-off. Therefore, before delving into the different types of model ensembles, it is necessary to review and understand the bias-variance trade-off principle (as it applies to the field of statistics or machine learning). In predictive analytics, *bias* refers to the error and *variance* refers to the consistency (or lack thereof) in predictive accuracy of models applied to other data sets. The best models are expected to have low bias (low error, high accuracy) and low variance (consistency of accuracy from data set to data set). Unfortunately, there is always a trade-off between these two metrics in building predictive models—improving one results in worsening the other. You can achieve low bias on training data, but the model could suffer from high variance on hold-out/validation data because the models could have been overtrained/overfit. For instance, the *k*NN algorithm with $k = 1$ is an example of a low bias model (perfect on training data set) but is susceptible to high variance on a test/validation data set. Use of cross-validation along with proper model ensembles seems to be the current best practice in handling such trade-offs between bias and variance in predictive modeling.

Different Types of Ensembles

Ensembles or teams of predictive models working together have been the fundamental strategy for developing accurate and robust analytics models. Although ensembles have been around for quite a while, their popularity and effectiveness has surfaced in a significant way only within the last decade as they continually improved in parallel with the rapidly improving software and hardware capabilities. When we refer to model ensembles, many of us immediately think of decision tree ensembles like random forest and boosted trees; however, generally speaking, the model ensembles can be classified into four groups in two dimensions as shown in Figure 5.20. The first dimension is the method type (the *x*-axis in Figure 5.20) in which the ensembles can be grouped into bagging or boosting types. The second dimension is the model type (the *y*-axis in Figure 5.20) in which the ensembles can be grouped into homogeneous or heterogeneous types (Abbott, 2014).

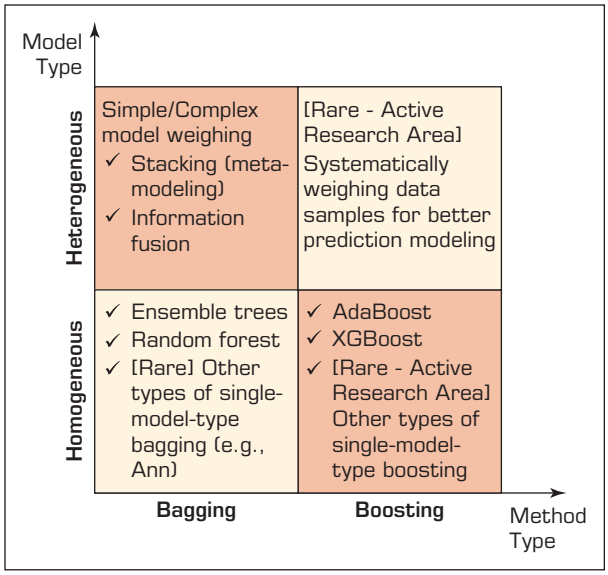


FIGURE 5.20 **Simple Taxonomy for Model Ensembles.**

As the name implies, *homogeneous-type ensembles* combine the outcomes of two or more of the same type of models such as decision trees. In fact, a vast majority of homogeneous model ensembles are developed using a combination of decision tree structures. The two most common categories of homogeneous type ensembles that use decision trees are bagging and boosting (more information on these are given in subsequent sections). Heterogeneous model ensembles combine the outcomes of two or more different types of models such as decision trees, artificial neural networks, logistic regression, SVM, and others. As mentioned in the context of "the wisdom of crowds," one of the key success factors in ensemble modeling is to use models that are fundamentally different from one another, ones that look at the data from a different perspective. Because of the way it combines the outcomes of different model types, heterogeneous model ensembles are also called *information fusion models* (Delen and Sharda, 2010) or stacking (more information on these is given later in this chapter).

Bagging

Bagging is the simplest and most common ensemble method. Leo Breiman, a very well-respected scholar in the world of statistics and analytics, is known to have first published a description of bagging (i.e., Bootstrap Aggregating) algorithm at the University of California–Berkeley in 1996 (Breiman, 1996). The idea behind bagging is quite simple yet powerful: build multiple decision trees from resampled data and combine the predicted values through averaging or voting. The resampling method Breiman used was bootstrap sampling (sampling with replacement), which creates replicates of some records in the training data. With this selection method, on average, about 37 percent of the records will not be included at all in the training data set (Abbott, 2014).

Although bagging was first developed for decision trees, the idea can be applied to any predictive modeling algorithm that produces outcomes with sufficient variation in the predicted values. Although rare in practice, the other predictive modeling algorithms that are potential candidates for bagging-type model ensembles include neural networks, Naïve Bayes, *k*-nearest neighbor (for low values of *k*), and, to a lesser degree, even logistic regression. *k*-nearest neighbor is not a good candidate for bagging if the value of *k* is

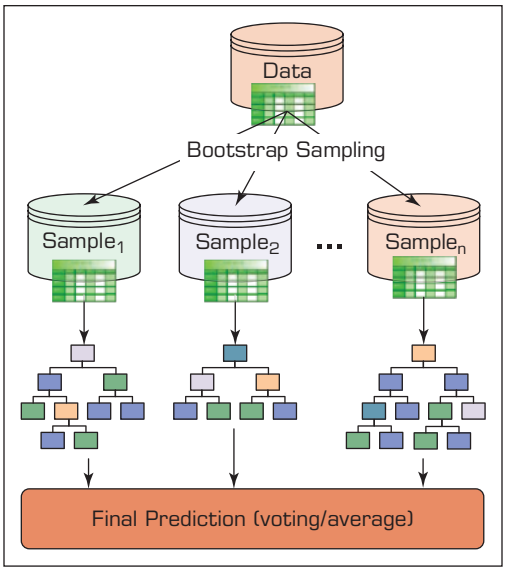


FIGURE 5.21 **Bagging-Type Decision Tree Ensembles.**

already large; the algorithm already votes or averages predictions and with larger values of *k*, so predictions are already very stable with low variance.

Bagging can be used for both classification- and regression/estimation–type prediction problems. In classification-type prediction problems, all of the participant models' outcomes (class assignments) are combined using either a simple or complex/weighted majority voting mechanism. The class label that gets the most/highest votes becomes the aggregated/ensemble prediction for that sample/record. In regression/estimation–type prediction problems, when the output/target variable is a number, all of the participant models' outcomes (numerical estimations) are combined using either a simple or complex/weighted–averaging mechanism. Figure 5.21 illustrates the graphical depiction of a decision tree–type bagging algorithm.

One of the key questions in bagging is, "How many bootstrap samples, also called *replicates*, should be created?" Brieman stated, "My sense of it is that fewer are required when *y* [dependent variable] is numerical and more are required with an increasing number of classes [for classification type prediction problems]." He typically used 10–25 bootstrap replicates with significant improvements occurring with as few as 10 replicates. Overfitting the models is an important requirement to building good bagged ensembles. By overfitting each model, the bias is low, but the decision tree generally has worse accuracy on held-out data. But bagging is a variance reduction technique; the averaging of predictions smooths the predictions to behave in a more stable way on new data.

As mentioned before, the diversity of model predictions is a key factor in creating effective ensembles. One way to measure the diversity of predictions is to examine the correlation of predicted values. If the correlations between model predictions are always very high, more than 0.95, each model brings little additional predictive information to the ensemble and therefore little improvement in accuracy is achievable. Generally, it is best to have correlations of less than 0.9. The correlations should be computed from the model propensities or predicted probabilities rather than the {0,1} classification value itself. Bootstrap sampling in bagging is the key to introducing diversity in the models. One can think of the bootstrap sampling methodology as creating case weights for each record—some records are included multiple times in the training data (their weights are

1, 2, 3, or more), and other records are not included at all (their weights are equal to 0) (Abbott, 2014).

Boosting

Boosting is perhaps the second most common ensemble method after bagging. Yoav Freund and Robert E. Schapire are known to have first introduced the boosting algorithm in separate publications in the early 1990s and then in a 1996 joint publication (Freund and Schapire, 1996). They introduced the well-known boosting algorithm, called AdaBoost. As with bagging, the idea behind boosting is also quite straightforward. First, build a rather simple classification model; it needs to be only slightly better than random chance, so for a binary classification problem, it needs to be only slightly better than a 50 percent correct classification. In this first step, each record is used in the algorithm with equal case weights as one would do normally in building a predictive model. The errors in the predicted values for each case are noted. The case weights of correctly classified records/cases/samples will stay the same or perhaps be reduced, and the case weights of the records that are incorrectly classified will have increased, and then a second simple model is built on these weighted cases (i.e., the transformed/weighted–training data set). In other words, for the second model, records that were incorrectly classified are "boosted" through case weights to be considered more strongly or seriously in the construction of the new prediction model. In each iteration, the records that are incorrectly predicted (the ones that are difficult to classify) keep having their case weights increased, communicating to the algorithm to pay more attention to these records until, hopefully, they are finally classified correctly.

This process of boosting is often repeated tens or even hundreds of times. After the tens or hundreds of iterations, the final predictions are made based on a weighted average of the predictions from all the models. Figure 5.22 illustrates the simple process of boosting in building decision tree–type ensemble models. As shown, each tree takes the most current data set (one of equal size, but with the most recently boosted case weights)

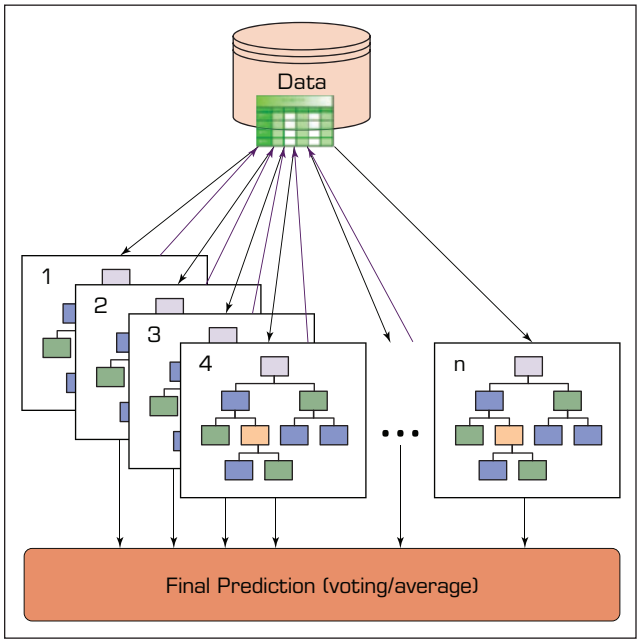


FIGURE 5.22 **Boosting-Type Ensembles for Decision Trees.**

to build another tree. The feedback of incorrectly predicted cases is used as an indicator to determine which cases and to what extent (direction and magnitude) to boost (update the weights) for the training samples/cases.

Although they look quite similar in structure and purpose, bagging and boosting employ slightly different strategies to utilize the training data set and to achieve the goal of building the best possible prediction model ensemble. The two key differences between bagging and boosting are as follows. Bagging uses a bootstrap sample of cases to build decision trees whereas boosting uses the complete training data set. Whereas bagging creates independent, simple trees to ensemble, boosting creates dependent trees (each tree "learning" from the previous one to pay more attention to the incorrectly predicted cases) that collectively contribute to the final ensemble.

Boosting methods are designed to work with weak learners, that is, simple models; the component models in a boosted ensemble are simple models with high bias although low variance. The improvement with boosting is better, as with bagging, when algorithms that are unstable predictors are used. Decision trees are most often used in boosted models. Naïve Bayes is also used but with fewer improvements over a single model. Empirically speaking, boosting typically produces better model accuracy than single decision trees or even bagging-type ensembles.

Variants of Bagging and Boosting

Bagging and boosting were the first ensemble methods that appeared in predictive analytics software, primarily with decision tree algorithms. Since their introduction, many other approaches to building ensembles have been developed and made available, particularly in open source software (both as part of open analytics platforms like KNIME and Orange and as class libraries in R and Python). The most popular and successful [advanced] variants of bagging and boosting are random forest and stochastic gradient boosting, respectively.

RANDOM FOREST The random forest (RF) model was first introduced by Breiman (2001) as a modification to the simple bagging algorithm. As with bagging, the RF algorithm begins with a bootstrap-sampled data set and builds one decision tree from each bootstrap sample. Compared to simple bagging, there is, however, an important twist to the RF algorithm: at each split in the tree, staring from the very first split, rather than considering all input variables as candidates, only a random subset of variables is considered. Hence, in RF, the bootstrap sampling technique applies to both a random selection of cases and a random selection of features (i.e., input variables).

The number of cases and the number of variables to consider along with how many trees to construct are all parameters used to decide in building RF models. Common practice suggests that the default number of variables to consider as candidates at each split point should be the square root of the total number of candidate inputs. For example, if there were 100 candidate inputs for the model, a random 10 inputs are candidates for each split. This also means that it is unlikely that the same inputs will be available for splits at parent and children nodes in a given tree, forcing the tree to find alternate ways to maximize the accuracy of subsequent splits. Therefore, there is an intentionally created twofold diversity mechanism built into the tree construction process—random selection of cases and variables. RF models produce prediction outcomes that are usually more accurate than simple bagging and are often more accurate than simple boosting (i.e., AdaBoost).

STOCHASTIC GRADIENT BOOSTING The simple boosting algorithm AdaBoost is only one of many boosting algorithms currently documented in the literature. In commercial software, AdaBoost is still the most commonly used boosting technique; however, dozens

of boosting variants can be found in open source software packages. One interesting boosting algorithm that has recently gained popularity due to its superior performance is the stochastic gradient boosting (SGB) algorithm created by Jerry Friedman at Stanford University. Then Friedman developed an advanced version of this algorithm (Friedman, 2001) called *multiple additive regression trees* (MART) and later branded as TreeNet® by Salford Systems in its software tool. Like other boosting algorithms, the MART algorithm builds successive, simple trees and combines them additively. Typically, the simple trees are more than stumps and contain up to six terminal nodes. Procedurally, after building the first tree, errors (also called *residuals*) are computed. The second tree and all subsequent trees then use residuals as the target variable. Subsequent trees identify patterns that relate inputs to small and large errors. Poor prediction of the errors results in large errors to predict in the next tree, and good predictions of the errors result in small errors to predict in the next tree. Typically, hundreds of trees are built, and the final predictions are additive combinations of the predictions that are, interestingly, piecewise constant models because each tree is itself a piecewise constant model. However, one rarely notices these intricacies about the individual trees because typically hundreds of trees are included in the ensemble (Abbott, 2014). The TreeNet algorithm, an example of stochastic gradient boosting, has won multiple data mining modeling competitions since its introduction and has proven to be an accurate predictor with the benefit that very little data cleanup is needed for the trees before modeling.

Stacking

Stacking (a.k.a. stacked generalization or super learner) is a part of heterogeneous ensemble methods. To some analytics professionals, it could be the optimum ensemble technique but is also the least understood (and the most difficult to explain). Due to its two-step model training procedure, some think of it as an overly complicated ensemble modeling. Simply put, stacking creates an ensemble from a diverse group of strong learners. In the process, it interjects a metadata step called *super learner* or *meta learner*. These intermediate meta classifiers forecast how accurate the primary classifiers have become and are used as the basis for adjustments and corrections (Vorhies, 2016). The process of stacking is figuratively illustrated in Figure 5.23.

As shown in Figure 5.23, in constructing a stacking-type model ensemble, a number of diverse strong classifiers are first trained using bootstrapped samples of the training data, creating tier 1 classifiers (each optimized to its full potential for the best possible prediction outcomes). The outputs of the tier 1 classifiers are then used to train a tier 2 classifier (i.e., a metaclassifier) (Wolpert, 1992). The underlying idea is to learn whether training data have been properly learned. For example, if a particular classifier incorrectly learned a certain region of the feature space and hence consistently misclassifies instances coming from that region, the tier 2 classifier might be able to learn this behavior and along with the learned behaviors of other classifiers, it can correct such improper training. Cross-validation–type selection is typically used for training the tier 1 classifiers—the entire training data set is divided into k numbers of mutually exclusive subsets, and each tier 1 classifier is first trained on (a different set of) $k - 1$ subsets of the training data. Each classifier is then evaluated on the k^{th} subset, which was not seen during training. The outputs of these classifiers on their pseudo-training blocks, along with the actual correct labels for those blocks, constitute the training data set for the tier 2 classifier.

Information Fusion

As part of heterogeneous model ensembles, information fusion combines (fuses) the output (i.e., predictions) of different types of models such as decision trees, artificial neural networks, logistic regression, SVM, Naïve Bayes, and k-nearest neighbor, among others,

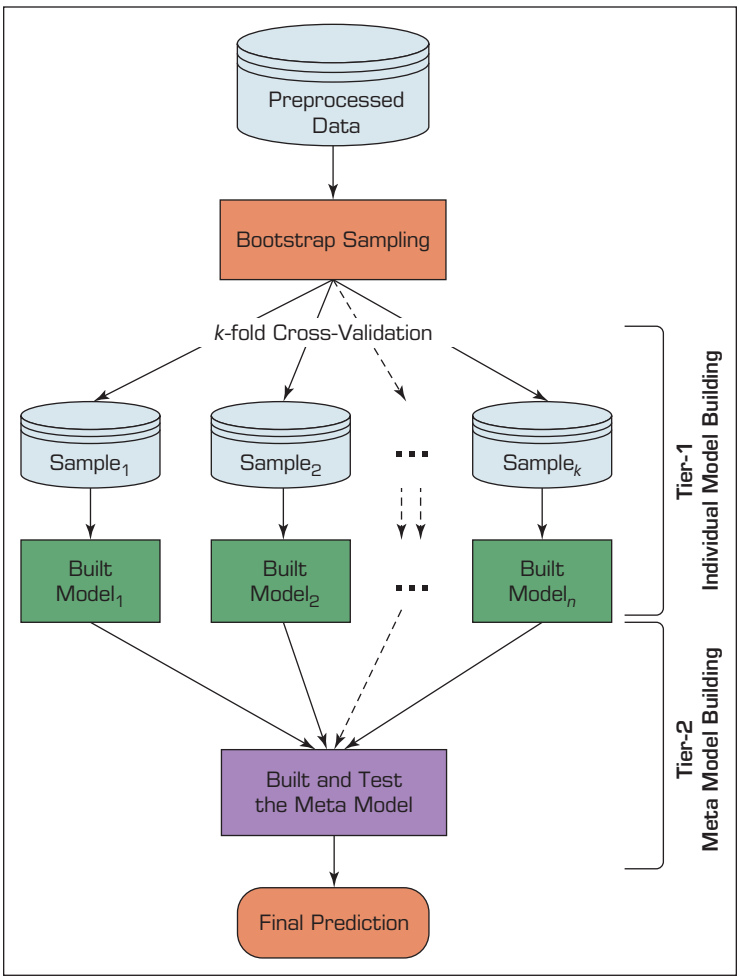


FIGURE 5.23 **Stacking-Type Model Ensembles.**

and their variants. The difference between stacking and information fusion is the fact that information fusion has no "meta modeling" or "superlearners." It simply combines the outcomes of heterogeneous strong classifiers using simple or weighted voting (for classification) or simple or weighted averaging (for regression). Therefore, it is simpler and less computationally demanding than stacking. In the process of combining the outcomes of multiple models, either a simple voting (each model contributes equally one vote) or a weighted combination of voting (each model contributes based on its prediction accuracy—more accurate models have higher weight value) can be used. Regardless of the combination method, this type of heterogeneous ensemble has been shown to be an invaluable addition to any data mining and predictive modeling project. Figure 5.24 graphically illustrates the process of building information fusion–type model ensembles.

Summary—Ensembles are not Perfect!

As a prospective data scientist, if you are asked to build a prediction model (or any other analytics model for that matter), you are expected to develop some of the popular model ensembles along with the standard individual models. If done properly, you will realize that ensembles are often more accurate and almost always more robust and reliable than

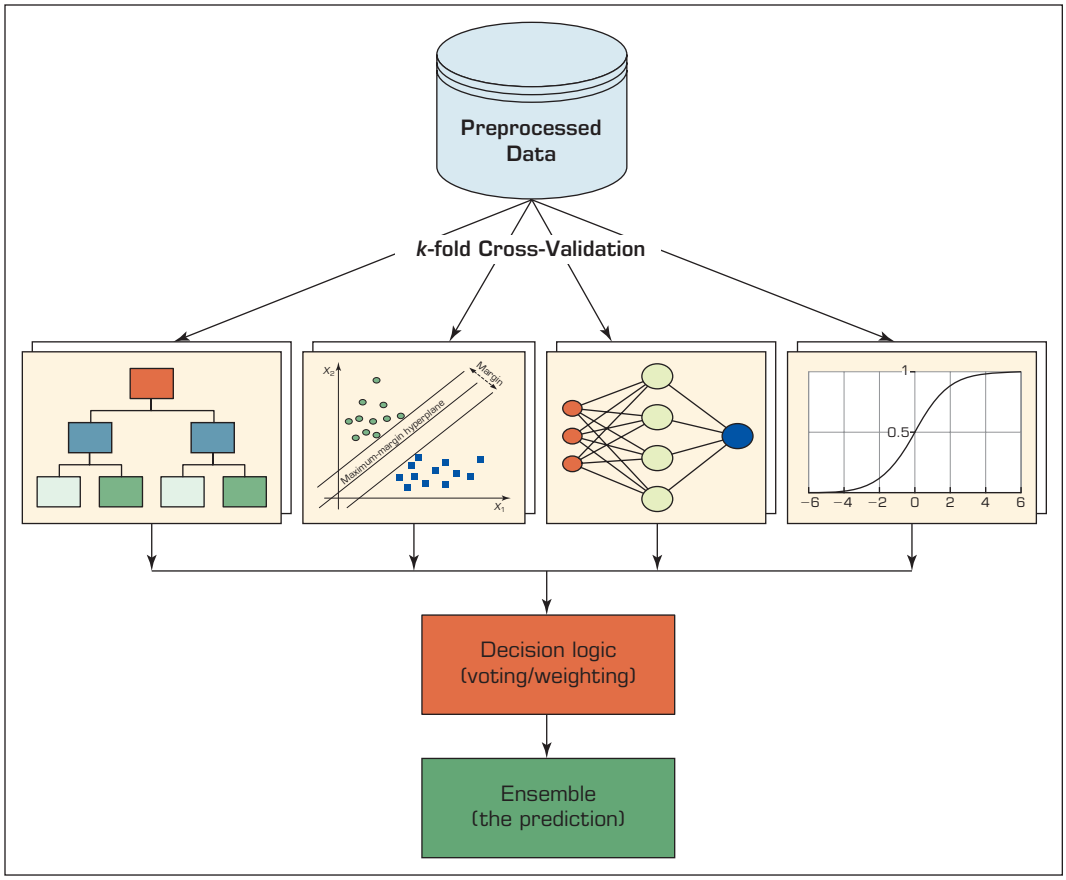


FIGURE 5.24 **Illustration of Building Process for Information Fusion–Type Model Ensemble.**

the individual models. Although they seem like silver bullets, model ensembles are not without shortcomings; the following are the two most common ones.

COMPLEXITY Model ensembles are more complex than individual models. Occam's razor is a core principle that many data scientists use; the idea is that simpler models are more likely to generalize better, so it is better to reduce/regularize complexity, or in other words, simplify the models so that the inclusion of each term, coefficient, or split in a model is justified by its power of reducing the error at a sufficient amount. One way to quantify the relationship between accuracy and complexity is taken from information theory in the form of information theoretic criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and minimum description length (MDL). Traditionally, statisticians—more recently, data scientists—are using these criteria to select variables in predictive modeling. Information theoretic criteria require a reduction in model error to justify additional model complexity. So the question is, "Do model ensembles violate Occam's razor?" Ensembles, after all, are much more complex than single models. According to Abbott (2014), if the ensemble accuracy is better on held-out data than single models, then the answer is "no" as long as we think about the complexity of a model in different terms—not just computational complexity but also behavioral complexity. Therefore, we should not fear that adding computational complexity (more terms, splits, or weights) will necessarily increase the complexity of models because sometimes the ensemble will significantly reduce the behavioral complexity.

TRANSPARENCY The interpretation of ensembles can become quite difficult. If you build an RF ensemble containing 200 trees, how do you describe why a prediction has a particular value? You can examine each of the trees individually, although this is clearly not practical. For this reason, ensembles are often considered black-box models, meaning that what they do is not transparent to the modeler or domain expert. Although you can look at the split statistics (which variables are more often picked to split early in those 200 trees) to artificially judge the level of contribution (a pseudo-variable of importance measure), each variable is contributing to the trained model ensemble. Compared to a single decision tree, such an investigation of 200 trees will be too difficult and not an intuitive way to interpret how the model comes up with a specific prediction. Another way to determine which inputs to the model are most important is to perform a sensitivity analysis.

In addition to complexity and transparency, model ensembles are also more difficult and computationally more expensive to build and much more difficult to deploy. Table 5.9 shows the pros and cons of ensemble models compared with individual models.

In summary, model ensembles are the new frontier for predictive modelers who are interested in accuracy by reducing either errors in the models or the risk that the models behave erratically. Evidence for this is clear from the dominance of ensembles in predictive analytics and data mining competitions: ensembles always win.

The good news for predictive modelers is that many techniques for building ensembles are built into software already. The most popular ensemble algorithms (bagging, boosting, stacking, and their variants) are available in nearly every commercial or open source software tool. Building customized ensembles is also supported in many software products whether based on a single algorithm or through a heterogeneous ensemble.

Ensembles are not appropriate for every solution—their applicability is determined by the modeling objectives defined during business understanding and problem definition—but they should be part of every predictive modeler's and data scientist's modeling arsenal.

TABLE 5.9 Brief List of Pros and Cons of Model Ensembles Compared to Individual Models

PROS (Advantages)	Description
✓ Accuracy	Model ensembles usually result in more accurate models than individual models.
✓ Robustness	Model ensembles tend to be more robust against outliers and noise in the data set than individual models.
✓ Reliability (stable)	Because of the variance reduction, model ensembles tend to produce more stable, reliable, and believable results than individual models.
✓ Coverage	Model ensembles tend to have a better coverage of the hidden complex patterns in the data set than individual models.
CONS (Shortcomings)	**Description**
✓ Complexity	Model ensembles are much more complex than individual models.
✓ Computationally expensive	Compared to individual models, ensembles require more time and computational power to build.
✓ Lack of transparency (explainability)	Because of their complexity, it is more difficult to understand the inner structure of model ensembles (how they do what they do) than individual models.
✓ Harder to deploy	Model ensembles are much more difficult to deploy in an analytics-based managerial decision-support system than single models.

Application Case 5.6

To Imprison or Not to Imprison: A Predictive Analytics-Based Decision Support System for Drug Courts

Introduction and Motivation

Analytics has been used by many businesses, organizations, and government agencies to learn from past experiences to more effectively and efficiently use their limited resources to achieve their goals and objectives. Despite all the promises of analytics, however, its multidimensional and multidisciplinary nature can sometimes disserve its proper, full-fledged application. This is particularly true for the use of predictive analytics in several social science disciplines because these domains are traditionally dominated by descriptive analytics (causal-explanatory statistical modeling) and might not have easy access to the set of skills required to build predictive analytics models. A review of the extant literature shows that drug court is one such area. While many researchers have studied this social phenomenon, its characteristics, its requirements, and its outcomes from a descriptive analytics perspective, there currently is a dearth of predictive analytics models that can accurately and appropriately predict who would (or would not) graduate from intervention and treatment programs. To fill this gap and to help authorities better manage the resources, and to improve the outcomes, this study sought to develop and compare several predictive analytics models (both single models and ensembles) to identify who would graduate from these treatment programs.

Ten years after President Richard Nixon first declared a "war on drugs," President Ronald Reagan signed an executive order leading to stricter drug enforcement, stating, "We're taking down the surrender flag that has flown over so many drug efforts; we are running up a battle flag." The reinforcement of the war on drugs resulted in an unprecedented 10-fold surge in the number of citizens incarcerated for drug offences during the following two decades. The skyrocketing number of drug cases inundated court dockets, overloaded the criminal justice system, and overcrowded prisons. The abundance of drug-related caseloads, aggravated by a longer processing time than that for most other felonies, imposed tremendous costs on state and federal departments of justice. Regarding the increased demand, court systems started to look for innovative ways to accelerate the inquest of drug-related cases. Perhaps analytics-driven decision support systems are the solution to the problem. To support this claim, the current study's goal was to build and compare several predictive models that use a large sample of data from drug courts across different locations to predict who is more likely to complete the treatment successfully. The researchers believed that this endeavor might reduce the costs to the criminal justice system and local communities.

Methodology

The methodology used in this research effort included a multi-step process that employed predictive analytics methods in a social science context. The first step of this process, which focused on understanding the problem domain and the need to conduct this study, was presented in the previous section. For the steps of the process, the researchers employed a structured and systematic approach to develop and evaluate a set of predictive models using a large and feature-rich real-world data set. The steps included data understanding, data preprocessing, model building, and model evaluation; they are reviewed in this section. The approach also involved multiple iterations of experimentations and numerous modifications to improve individual tasks and to optimize the modeling parameters to achieve the best possible outcomes. A pictorial depiction of the methodology is given in Figure 5.25.

The Results

A summary of the models' performances based on accuracy, sensitivity, specificity, and AUC is presented in Table 5.10. As the results show, RF has the best classification accuracy and the greatest AUC among the models. The heterogeneous ensemble_ (HE) model closely follows RF, and SVM, ANN, and LR rank third to last based on their classification performances. RF also has the highest specificity and the second highest sensitivity. Sensitivity in the context of this study is an indicator of a model's ability in correctly predicting the outcome for successfully graduated participants. Specificity, on the other hand, determines how a model performs in predicting the end results for those who do not successfully complete the treatment. Consequently, it can be concluded that RF outperforms other models for the drug courts data set used in this study.

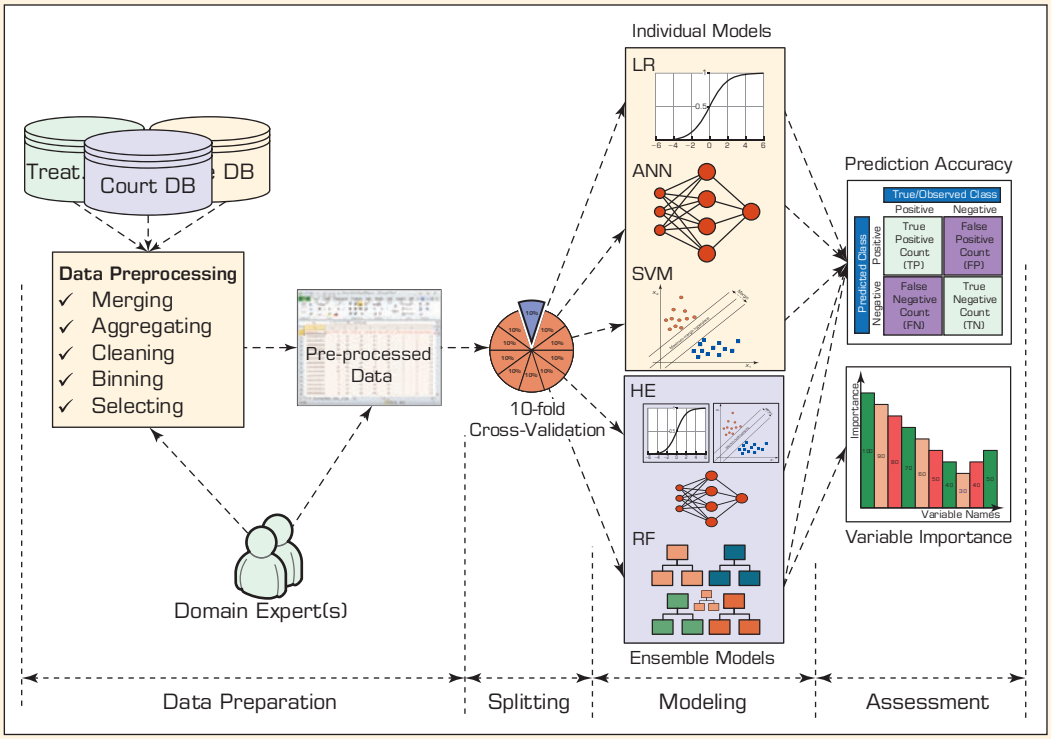


FIGURE 5.25 **Research Methodology Depicted as a Workflow.**

TABLE 5.10 **Performance of Predictive Models Using 10-Fold Cross-Validation on the Balanced Data Set**

Model Type			Confusion Matrix G	Confusion Matrix T	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Individual Models	ANN	G	6,831	1,072	86.63	86.76	86.49	0.909
		T	1,042	6,861				
	SVM	G	6,911	992	88.67	89.63	87.75	0.917
		T	799	7,104				
	LR	G	6,321	1,582	85.13	86.16	81.85	0.859
		T	768	7,135				
Ensembles	RF	G	6,998	905	91.16	93.44	89.12	0.927
		T	491	7,412				
	HE	G	6,885	1,018	90.61	93.66	87.96	0.916
		T	466	7,437				

ANN: artificial neural networks; DT: decision trees; LR: logistic regression; RF: random forest; HE: heterogeneous ensemble; AUC: area under the curve; G: graduated; T: terminated

(*Continued*)

Application Case 5.6 (Continued)

Although the RF model performs better than the other models in general, it falls second to the HE model in the number of false negative predictions. Similarly, the HE model has a slightly better performance in true negative predictions. False positive predictions represent participants who were terminated from the treatment, but the models mistakenly classified them as successful graduates. False negatives pertain to individuals who graduated, but the models predicted them as dropouts. False positive predictions are synonymous with increased costs and opportunity losses whereas false negatives carry social impacts. Spending resources on those offenders who would recidivate at some point in time during the treatment and, hence, be terminated from the program prevented a number of (potentially successful) prospective offenders from participating in the treatment. Conspicuously, depriving potentially successful offenders from the treatment is against the initial objective of drug courts in reintegrating nonviolent offenders into their communities.

In summary, traditional causal-explanatory statistical modeling, or descriptive analytics, uses statistical inference and significance levels to test and evaluate the explanatory power of hypothesized underlying models or to investigate the association between variables retrospectively. Although a legitimate approach for understanding the relationships within the data used to build the model, descriptive analytics falls short in predicting outcomes for prospective observations. In other words, partial explanatory power does not imply predictive power, and predictive analytics is a must for building empirical models that predict well. Therefore, relying on the findings of this study, application of predictive analytics (rather than the sole use of descriptive analytics) to predict the outcomes of drug courts is well grounded.

Questions for Case 5.6

1. What are drug courts and what do they do for the society?
2. What are the commonalities and differences between traditional (theoretical) and modern (machine-learning) base methods in studying drug courts?
3. Can you think of other social situations and systems for which predictive analytics can be used?

Source: Zolbanin, H., and Delen, D. (2018). To Imprison or Not to Imprison: An Analytics-Based Decision Support System for Drug Courts. *The Journal of Business Analytics (forthcoming).*

SECTION 5.9 REVIEW QUESTIONS

1. What is a model ensemble, and where can it be used analytically?
2. What are the different types of model ensembles?
3. Why are ensembles gaining popularity over all other machine-learning trends?
4. What is the difference between bagging- and boosting-type ensemble models?
5. What are the advantages and disadvantages of ensemble models?

Chapter Highlights

- Neural computing involves a set of methods that emulates the way the human brain works. The basic processing unit is a neuron. Multiple neurons are grouped into layers and linked together.
- There are differences between biological and artificial neural networks.
- In an artificial neural network, knowledge is stored in the weight associated with each connection between two neurons.
- Neural network applications abound in almost all business disciplines as well as in virtually all other functional areas.
- Business applications of neural networks include finance, bankruptcy prediction, time-series forecasting, and so on.
- There are various neural network architectures for different types of problems.
- Neural network architectures can be applied not only to prediction (classification or estimation)

but also to clustering and optimization-type problems.
- SVM are among popular machine-learning techniques, mostly because of their superior predictive performance and their theoretical foundation.
- Although SVM can use a radial-basis function as a kernel, they are not very similar to neural networks.
- SVM can be used for both classification- and estimation/regression-type prediction problems.
- SVM use only numerical variables and the supervised machine-learning method.
- Plenty of SVM applications exist, and new ones are emerging in a variety of domains including healthcare, finance, security, and energy.
- The nearest neighbor (or *k*-nearest neighbor) algorithm is a simple machine-learning technique that is used for both classification- and estimation/regression-type prediction problems.
- The nearest neighbor algorithm is a type of instance-based learning (or lazy learning) algorithm in which all computations are deferred until the actual prediction.
- The parameter *k* signifies the number of neighbors to use in a given prediction problem.
- Determining the "optimal" value of *k* requires a cross-validation–type experimentation.
- The nearest neighbor algorithm uses a distance measure to identify close-by/appropriate neighbors.
- The input variables to the nearest neighbor algorithm must be in numeric format; all nonnumeric/nominal variables need to be converted to pseudo-binary numeric variables.
- Bayesian classifiers are built on the foundation of the Bayes theorem (i.e., conditional probabilities).
- Naïve Bayes is a simple probability-based classification method that is applied to classification-type prediction problems.
- The Naïve Bayes method requires input and output variables to have nominal values; numeric ones need to be discretized.
- *Naïve keyword* refers to the somewhat unrealistic yet practical assumption of independence (of the predictor/input variables).
- The Bayesian network (or Bayesian belief network) is a relatively new machine-learning technique that is gaining popularity among data scientists, academics, and theorists.
- The Bayesian network is a powerful tool for representing dependency structure in a graphical, explicit, and intuitive way.
- The Bayesian network can be used for prediction and explanation (or the interrelationships among the variables).
- Bayesian networks can be constructed manually (based on a domain expert's knowledge) or automatically using historical data.
- While constructing a Bayesian network automatically, one can use regular Naïve Bayes or the tree-augmented Naïve (TAN) Bayes.
- Bayesian networks provide an excellent model for conducting what-if analyses for a variety of hypothetical scenarios.
- Ensembles (or more appropriately *model ensembles* or *ensemble modeling*) are combinations of the outcomes produced by two or more analytics models into a compound output.
- Although ensembles are primarily used for prediction modeling when the scores of two or more models are combined to produce a better prediction, they can also be used for clustering and association.
- Ensembles can be applied to both classification (via voting) and estimation/regression-type (via averaging) prediction problems.
- Ensembles are used mainly for two reasons: to obtain better accuracy and achieve more stable/reliable outcomes.
- Recent history in data science has shown that ensembles win competitions.
- There are homogeneous and heterogeneous ensembles; if the combined models are of the same type (e.g., decision trees), the ensemble is homogeneous; if not, it is heterogeneous.
- There are three methods in ensemble modeling: bagging, boosting, and stacking.
- Random forest is a bagging-type, homogeneous, decision tree–based ensemble method.
- Stochastic gradient boosting is a boosting type that is a homogeneous, decision tree–based ensemble method.
- Information fusion and stacking are heterogeneous ensembles in which different types of models are combined.
- The disadvantages of ensembles include complexity and lack of transparency.

Key Terms

AdaBoost
artificial neural network (ANN)
attrition
axon
backpropagation
bagging
Bayesian belief network (BNN)
Bayesian network (BN)
Bayes theorem
boosting
conditional probability
cross-validation
dendrites
distance metric
Euclidean distance
heterogeneous ensemble
hidden layer
Hopfield network
hyperplane
information fusion
k-fold cross-validation
k-nearest neighbor (*k*NN)
kernel trick
Kohonen's self-organizing feature map
Manhattan distance
maximum margin
Minkowski distance
multi-layer perceptron
Naïve Bayes
neural computing
neural network
neuron
nucleus
pattern recognition
perceptron
processing element (PE)
radial basis function (RBF)
random forest
retention
stacking
supervised learning
stochastic gradient boosting
synapse
transformation (transfer) function
voting
weights
what-if scenario

Questions for Discussion

1. What is an artificial neural network and for what types of problems can it be used?
2. Compare artificial and biological neural networks. What aspects of biological networks are not mimicked by artificial ones? What aspects are similar?
3. What are the most common ANN architectures? For what types of problems can they be used?
4. ANN can be used for both supervised and unsupervised learning. Explain how they learn in a supervised mode and in an unsupervised mode.
5. What are SVM? How do they work?
6. What are the types of problems that can be solved by SVM?
7. What is the meaning of "maximum-margin hyperplanes"? Why are they important in SVM?
8. What is the kernel trick and how does it relate to SVM?
9. What are the specific steps to follow in developing an SVM model?
10. How can the optimal kernel type and kernel parameters be determined?
11. What are the common application areas for SVM? Conduct a search on the Internet to identify popular application areas and specific SVM software tools used in those applications.
12. What are the commonalities and differences, advantaged and disadvantages between ANN and SVM?
13. Explain the difference between a training and a testing data set in ANN and SVM. Why do we need to differentiate them? Can the same set be used for both purposes? Why or why not?
14. Everyone would like to make a great deal of money on the stock market. Only a few are very successful. Why is using an SVM or ANN a promising approach? What can they do that other decision support technologies cannot do? How could SVM or ANN fail?
15. What is special about the *k*NN algorithm?
16. What are the advantages and disadvantages of *k*NN as compared to ANN and SVM?
17. What are the critical success factors for a *k*NN implementation?
18. What is a similarity (or distance) measure? How can it be applied to both numerical and nominal valued variables?
19. What are the common (business and scientific) applications of *k*NN? Conduct a Web search to find three real-world applications that use *k*NN to solve the problem.
20. What is special about the Naïve Bayes algorithm? What is the meaning of "Naïve" in this algorithm?
21. What are the advantages and disadvantages of Naïve Bayes compared to other machine-learning methods?
22. What type of data can be used in a Naïve Bayes algorithm? What type of predictions can be obtained from it?
23. What is the process of developing and testing a Naïve Bayes classifier?
24. What are Bayesian networks? What is special about them?
25. What is the relationship between Naïve Bayes and Bayesian networks?
26. What is the process of developing a Bayesian networks model?
27. What are the advantages and disadvantages of Bayesian networks compared to other machine-learning methods?
28. What is Tree Augmented Naïve (TAN) Bayes and how does it relate to Bayesian networks?
29. What is a model ensemble, and analytically where can it be used?
30. What are the different types of model ensembles?
31. Why are ensembles gaining popularity over all other machine-learning trends?
32. What is the difference between bagging- and boosting-type ensemble models?
33. What are the advantages and disadvantages of ensemble models?

Exercises

Teradata University Network (TUN) and Other Hands-On Exercises

1. Go to the Teradata University Network Web site (**teradatauniversitynetwork.com**) or a URL given by your instructor. Locate Web seminars related to data mining and neural networks. Specifically, view the seminar given by Professor Hugh Watson at the SPIRIT2005 conference at Oklahoma State University; then, answer the following questions:
   a. Which real-time application at Continental Airlines might have used a neural network?
   b. What inputs and outputs can be used in building a neural network application?
   c. Given that its data mining applications are in real time, how might Continental implement a neural network in practice?
   d. What other neural network applications would you propose for the airline industry?
2. Go to the Teradata University Network Web site (**teradatauniversitynetwork.com**) or a URL given by your instructor. Locate the Harrah's case. Read the case and answer the following questions:
   a. Which of the Harrah's data applications are most likely implemented using neural networks?
   b. What other applications could Harrah's develop using the data it collects from its customers?
   c. What are some concerns you might have as a customer at this casino?
3. A bankruptcy-prediction problem can be viewed as a problem of classification. The data set you will be using for this problem includes five ratios that have been computed from the financial statements of real-world firms. These five ratios have been used in studies involving bankruptcy prediction. The first sample includes data on firms that went bankrupt and firms that did not. This will be your training sample for the neural network. The second sample of 10 firms also consists of some bankrupt firms and some nonbankrupt firms. Your goal is to use neural networks, SVM, and nearest neighbor algorithms to build a model using the first 20 data points and then to test its performance on the other 10 data points. (Try to analyze the new cases yourself manually before you run the neural network and see how well you do.) The following tables show the training sample and test data you should use for this exercise.

Training Sample						
Firm	WC/TA	RE/TA	EBIT/TA	MVE/TD	S/TA	BR/NB
1	0.1650	0.1192	0.2035	0.8130	1.6702	1
2	0.1415	0.3868	0.0681	0.5755	1.0579	1
3	0.5804	0.3331	0.0810	1.1964	1.3572	1
4	0.2304	0.2960	0.1225	0.4102	3.0809	1
5	0.3684	0.3913	0.0524	0.1658	1.1533	1
6	0.1527	0.3344	0.0783	0.7736	1.5046	1
7	0.1126	0.3071	0.0839	1.3429	1.5736	1
8	0.0141	0.2366	0.0905	0.5863	1.4651	1
9	0.2220	0.1797	0.1526	0.3459	1.7237	1
10	0.2776	0.2567	0.1642	0.2968	1.8904	1
11	0.2689	0.1729	0.0287	0.1224	0.9277	0
12	0.2039	−0.0476	0.1263	0.8965	1.0457	0
13	0.5056	−0.1951	0.2026	0.5380	1.9514	0
14	0.1759	0.1343	0.0946	0.1955	1.9218	0
15	0.3579	0.1515	0.0812	0.1991	1.4582	0
16	0.2845	0.2038	0.0171	0.3357	1.3258	0
17	0.1209	0.2823	−0.0113	0.3157	2.3219	0
18	0.1254	0.1956	0.0079	0.2073	1.4890	0
19	0.1777	0.0891	0.0695	0.1924	1.6871	0
20	0.2409	0.1660	0.0746	0.2516	1.8524	0

Test Data						
Firm	WC/TA	RE/TA	EBIT/TA	MVE/TD	S/TA	BR/NB
A	0.1759	0.1343	0.0946	0.1955	1.9218	?
B	0.3732	0.3483	−0.0013	0.3483	1.8223	?
C	0.1725	0.3238	0.1040	0.8847	0.5576	?
D	0.1630	0.3555	0.0110	0.3730	2.8307	?
E	0.1904	0.2011	0.1329	0.5580	1.6623	?
F	0.1123	0.2288	0.0100	0.1884	2.7186	?
G	0.0732	0.3526	0.0587	0.2349	1.7432	?
H	0.2653	0.2683	0.0235	0.5118	1.8350	?
I	0.1070	0.0787	0.0433	0.1083	1.2051	?
J	0.2921	0.2390	0.0673	0.3402	0.9277	

Describe the results of the neural network, SVM, and nearest neighbor model predictions, including software, architecture, and training information.

4. The purpose of this exercise is to develop models to predict the type of forest cover using a number of cartographic measures. The given data set (see Online Supplements) includes four wilderness areas found in the Roosevelt National Forest of northern Colorado. A total of 12 cartographic measures were utilized as independent variables; seven major forest cover types were used as dependent variables.

   This is an excellent example for a multi-class classification problem. The data set is rather large (with 581,012 unique instances) and feature rich. As you will see, the data are also raw and skewed (unbalanced for different cover types). As a model builder, you are to make necessary decisions to preprocess the data and build the best possible predictor. Use your favorite tool to build the models for neural networks, SVM, and nearest neighbor algorithms, and document the details of your results and experiences in a written report. Use screenshots within your report to illustrate important and interesting findings. You are expected to discuss and justify any decision that you make along the way.

5. Go to UCI Machine-Learning Repository (**archive.ics.uci.edu/ml/index.php**), identify four data sets for classification-type problems, and use these data sets to build and compare ANN, SVM, *k*NN, and Naïve Bayes models. To do so, you can use any analytics tool. We suggest you use a free, open-source analytics tool such as KNIME (**knime.org**) or Orange (**orange.biolab.si**). Prepare a well-written report to summarize your findings.

6. Go to Google Scholar (**scholar.google.com**). Conduct a search to find two papers written in the last five years that compare and contrast multiple machine-learning methods for a given problem domain. Observe commonalities and differences among their findings and prepare a report to summarize your understanding.

Team Assignments and Role-Playing Projects

1. Consider the following set of data that relates daily electricity usage to a function of the outside high temperature (for the day):

Temperature, *X*	Kilowatts, *Y*
46.8	12,530
52.1	10,800
55.1	10,180
59.2	9,730
61.9	9,750
66.2	10,230
69.9	11,160
76.8	13,910
79.7	15,110
79.3	15,690
80.2	17,020
83.3	17,880

Number	Name	Independent Variables
1	Elevation	Elevation in meters
2	Aspect	Aspect in degrees azimuth
3	Slope	Slope in degrees
4	Horizontal_Distance_To_Hydrology	Horizontal distance to nearest surface water features
5	Vertical_Distance_To_Hydrology	Vertical distance to nearest surface water features
6	Horizontal_Distance_To_Roadways	Horizontal distance to nearest roadway
7	Hillshade_9am	Hill shade index at 9 A.M., summer solstice
8	Hillshade_Noon	Hill shade index at noon, summer solstice
9	Hillshade_3pm	Hill shade index at 3 P.M., summer solstice
10	Horizontal_Distance_To_Fire_Points	Horizontal distance to nearest wildfire ignition points
11	Wilderness_Area (4 binary variables)	Wilderness area designation
12	Soil_Type (40 binary variables)	Soil-type designation
Number	**Dependent Variable**	
1	Cover_Type (7 unique types)	Forest cover–type designation

Note: More details about the data set (variables and observations) can be found in the online file.

a. Plot the raw data. What pattern do you see? What do you think is really affecting electricity usage?
b. Solve this problem with linear regression $Y = a + bX$ (in a spreadsheet). How well does this work? Plot your results. What is wrong? Calculate the sum-of-the-squares error and R^2.
c. Solve this problem by using nonlinear regression. We recommend a quadratic function, $Y = a + b_1X + b_2X^2$. How well does this work? Plot your results. Is anything wrong? Calculate the sum-of-squares error and R^2.
d. Break the problem into three sections (look at the plot). Solve it using three linear regression models, one for each section. How well does this work? Plot your results. Calculate the sum-of-squares error and R^2. Is this modeling approach appropriate? Why or why not?
e. Build a neural network to solve the original problem. (You might have to scale the X and Y values to be between 0 and 1.) Train the network (on the entire set of data) and solve the problem (i.e., make predictions for each of the original data items). How well does this work? Plot your results. Calculate the sum-of-squares error and R^2.
f. Which method works best and why?

2. Build a real-world neural network. Using demo software downloaded from the Web (e.g., NeuroSolutions at **neurodimension.com** or another neural network tool/site), identify real-world data (e.g., start searching on the Web at **archive.ics.uci.edu/ml/index.php** or use data from an organization with which someone in your group has a contact) and build a neural network to make predictions. Topics might include sales forecasts, predicting success in an academic program (e.g., predict GPA from high school ranking and SAT scores, being careful to look out for "bad" data, such as GPAs of 0.0) or housing prices; or survey the class for weight, gender, and height and try to predict height based on the other two factors. You could also use U.S. Census data by state on this book's Web site or at **census.gov** to identify a relationship between education level and income. How good are your predictions? Compare the results to predictions generated using standard statistical methods (regression). Which method is better? How could your system be embedded in a decision support system (DSS) for real decision making?

3. For each of the following applications, would it be better to use a neural network or an expert system? Explain your answers, including possible exceptions or special conditions.
a. Diagnosis of a well-established but complex disease
b. Price lookup subsystem for a high-volume merchandise seller
c. Automated voice inquiry processing system
d. Training of new employees
e. Handwriting recognition

4. Consider the following data set, which includes three attributes and a classification for admission decisions into an MBA program:

GMAT	GPA	Quantitative GMAT	Decision
650	2.75	35	NO
580	3.50	70	NO
600	3.50	75	YES
450	2.95	80	NO
700	3.25	90	YES
590	3.50	80	YES
400	3.85	45	NO
640	3.50	75	YES
540	3.00	60	?
690	2.85	80	?
490	4.00	65	?

a. Using the data given here as examples, develop your own manual expert rules for decision making.
b. Build and test a neural network model using your favorite data mining tool. Experiment with different model parameters to "optimize" the predictive power of your model.
c. Build and test a support vector machine model using your favorite data mining tool. Experiment with different model parameters to "optimize" your model's predictive power. Compare the results with ANN and SVM.
d. Report the predictions on the last three observations from each of the three classification approaches (ANN, SVM, and *k*NN). Comment on the results.
e. Comment on the similarity and differences of these three prediction approaches. What did you learn from this exercise?

5. You have worked on neural networks and other data mining techniques. Give examples of the use of each of them. Based on your knowledge, how would you differentiate among these techniques? Assume that a few years from now you will come across a situation in which neural network or other data mining techniques could be used to build an interesting application for your organization. You have an intern working with you to do the grunt work. How will you decide whether the application is appropriate for a neural network or another data mining model? Based on your homework assignments, what specific software guidance can you provide so that your intern is productive for you quickly? Your answer for this question might mention the specific software, describe how to go about setting up the model/neural network, and validate the application.

Internet Exercises

1. Explore the Web sites of several neural network vendors, such as California Scientific Software (**calsci.com**), NeuralWare (**neuralware.com**), and Ward Systems Group (**wardsystems.comv**), and review some of their products. Download at least two demos and install, run, and compare them.

2. A very good repository of data that have been used to test the performance of neural network and other machine-learning algorithms can be accessed at **https://archive.ics.uci.edu/ml/index.php**. Some of the data sets are really meant to test the limits of current machine-learning algorithms and compare their performance against new approaches to learning. However, some of the smaller data sets can be useful for exploring the functionality of the software you might download in Internet Exercise 1 or the software that is available at **StatSoft.com** (i.e., Statistica Data Miner with extensive neural network capabilities). Download at least one data set from the UCI repository (e.g., Credit Screening Databases, Housing Database). Then apply neural networks as well as decision tree methods as appropriate. Prepare a report on your results. (Some of these exercises could also be completed in a group or even as semester-long projects for term papers and so on.)

3. Go to **calsci.com** and read about the company's various business applications. Prepare a report that summarizes the applications.

4. Go to **nd.com**. Read about the company's applications in investment and trading. Prepare a report about them.

5. Go to **nd.com**. Download the trial version of NeuroSolutions for Excel and experiment with it using one of the data sets from the exercises in this chapter. Prepare a report about your experience with the tool.

6. Go to **neoxi.com**. Identify at least two software tools that have not been mentioned in this chapter. Visit Web sites of those tools and prepare a brief report on their capabilities.

7. Go to **neuroshell.com**. Look at Gee Whiz examples. Comment on the feasibility of achieving the results claimed by the developers of this neural network model.

8. Go to **easynn.com**. Download the trial version of the software. After the installation of the software, find the sample file called **Houseprices.tvq**. Retrain the neural network and test the model by supplying some data. Prepare a report about your experience with this software.

9. Visit **tibco.com**. Download at least three white papers of applications. Which of these applications might have used neural networks?

10. Go to **neuralware.com**. Prepare a report about the products the company offers.
11. Go to **ibm.com**. Download at least two customer success stories or case studies that use advanced analytics or machine learning. Prepare a presentation for your understanding of these application cases.
12. Go to **sas.com**. Download at least two customer success stories or case studies that use advanced analytics or machine learning. Prepare a presentation for your understanding of these application cases.
13. Go to **teradata.com**. Download at least two customer success stories or case studies where advanced analytics or machine learning is used. Prepare a presentation for your understanding of these application cases.

References

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Hoboken, NJ: John Wiley.

Aizerman, M., E. Braverman, & L. Rozonoer. (1964). "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning." *Automation and Remote Control*, Vol. 25, pp. 821–837.

American Heart Association, "Heart Disease and Stroke Statistics," **heart.org** (accessed May 2018).

Boiman, E. S., & M. Irani. (2008). "In Defense of Nearest-Neighbor Based Image Classification," *IEEE Conference on Computer Vision and Pattern Recognition, 2008 (CVPR)*, 2008, pp. 1–8.

Bouzembrak, Y., & H. J. Marvin. (2016). "Prediction of Food Fraud Type Using Data from Rapid Alert System for Food and Feed (RASFF) and Bayesian Network Modelling." *Food Control, 61*, 180–187.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (2001). "Random Forests." *Machine Learning, 45* (1), 5–32.

Chow, C., & C. Liu (1968). "Approximating Discrete Probability Distributions with Dependence Trees." *IEEE Transactions on Information Theory, 14*(3), 462–473.

Delen, D., & R. Sharda. (2010). "Predicting the Financial Success of Hollywood Movies Using an Information Fusion Approach." *Indus Eng J, 21* (1), 30–37.

Delen, D., L. Tomak, K. Topuz, & E. Eryarsoy (2017). Investigating Injury Severity Risk Factors in Automobile Crashes with Predictive Analytics and Sensitivity Analysis Methods. *Journal of Transport & Health, 4*, 118–131.

Delen, D., A. Oztekin, & L. Tomak. (2012). "An Analytic Approach to Better Understanding and Management of Coronary Surgeries." *Decision Support Systems, 52* (3), 698–705.

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189–1232.

Freund, Y., & R. E. Schapire. (1996, July). "Experiments with a New Boosting Algorithm." In *Icml* (Vol. 96, pp. 148–156).

Friedman, D. (2014). "Oral Testimony before the House Committee on Energy and Commerce, by the Subcommittee on Oversight and Investigations," April 1, 2014, **www.nhtsa.gov/Testimony** (accessed October 2014).

Friedman, N., D. Geiger, & M. Goldszmidt. (1997). "Bayesian Network Classifiers." *Machine Learning*, Vol. 29, No. 2–3, 131–163.

Haykin, S. (2009). *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Hopfield, J. (1982, April). "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proceedings of National Academy of Science*, Vol. 79, No. 8, 2554–2558.

Koller, D., & N. Friedman. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Boston, MA: MIT Press.

McCulloch, W., & W. Pitts. (1943). "A Logical Calculus of the Ideas Imminent in Nervous Activity." *Bulletin of Mathematical Biophysics*, Vol. 5.

Medsker, L., & J. Liebowitz. (1994). *Design and Development of Expert Systems and Neural Networks*. New York, NY: Macmillan, p. 163.

Meyfroidt, G., F. Güiza, J. Ramon, & M. Bruynooghe. (2009). "Machine Learning Techniques to Examine Large Patient Databases." *Best Practice & Research Clinical Anaesthesiology, 23*(1), 127–143.

Minsky, M., & S. Papert. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Neural Technologies. "Combating Fraud: How a Leading Telecom Company Solved a Growing Problem." **neuralt.com/iqs/dlsfa.list/dlcpti.7/downloads.html** (accessed March 2009).

NHTSA (2018) National Highway Traffic Safety Administration (NHTSA's) General Estimate System (GES), **www.nhtsa.gov** (accessed January 20, 2017).

Pearl, J. (1985). "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning." *Proceedings of the Seventh Conference of the Cognitive Science Society*, 1985, pp. 329–334.

Pearl, J. (2009). *Causality*. Cambridge University Press, Cambridge, England.

Principe, J., N. Euliano, & W. Lefebvre. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. New York, NY: Wiley.

Reddy, B. K., D. Delen, & R. K. Agrawal. (2018). "Predicting and Explaining Inflammation in Crohn's Disease Patients Using Predictive Analytics Methods and Electronic Medical Record Data." *Health Informatics Journal*, 1460458217751015.

Reagan, R. (1982). "Remarks on Signing Executive Order 12368, Concerning Federal Drug Abuse Policy Functions," June 24, 1982. Online by Gerhard Peters and John

T. Woolley, The American Presidency Project. **http://www.presidency.ucsb.edu/ws/?pid=42671**.

Surowiecki, J. (2006). *The Wisdom of Crowds*. New York, NY: Penguin Random House.

Topuz, K., F. Zengul, A. Dag, A. Almehmi, & M. Yildirim (2018). "Predicting Graft Survival Among Kidney Transplant Recipients: A Bayesian Decision Support Model." *Decision Support Systems, 106*, 97–109.

Vorhies, W. (2016). "Want to Win Competitions? Pay Attention to Your Ensembles." Data Science Central Web Portal, **www.datasciencecentral.com/profiles/blogs/want-to-win-at-kaggle-pay-attention-to-your-ensembles** (accessed July 2018).

Wang, G., T. Xu, T. Tang, T. Yuan, & H. Wang. (2017). A "Bayesian Network Model for Prediction of Weather-Related Failures in Railway Turnout Systems." *Expert Systems with Applications, 69*, 247–256.

Wolpert, D. (1992). "Stacked Generalization." *Neural Networks, 5(2)*, 241–260.

Zahedi, F. (1993). *Intelligent Systems for Business: Expert Systems with Neural Networks*, Wadsworth, Belmont, CA.

Zhang, H., A. C. Berg, M. Maire, & J. Malik. (2006). "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition." In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* (Vol. 2, pp. 2126–2136). IEEE.

CHAPTER 6

Deep Learning and Cognitive Computing

LEARNING OBJECTIVES

- Learn what deep learning is and how it is changing the world of computing
- Know the placement of deep learning within the broad family of artificial intelligence (AI) learning methods
- Understand how traditional "shallow" artificial neural networks (ANN) work
- Become familiar with the development and learning processes of ANN
- Develop an understanding of the methods to shed light into the ANN black box
- Know the underlying concept and methods for deep neural networks
- Become familiar with different types of deep learning methods
- Understand how convolutional neural networks (CNN) work
- Learn how recurrent neural networks (RNN) and long short-memory networks (LSTM) work
- Become familiar with the computer frameworks for implementing deep learning
- Know the foundational details about cognitive computing
- Learn how IBM Watson works and what types of application it can be used for

Artificial intelligence (AI) is making a re-entrance into the world of commuting and in our lives, this time far stronger and much more promising than before. This unprecedented re-emergence and the new level of expectations can largely be attributed to deep learning and cognitive computing. These two latest buzzwords define the leading edge of AI and machine learning today. Evolving out of the traditional artificial neural networks (ANN), deep learning is changing the very foundation of how machine learning works. Thanks to large collections of data and improved computational resources, deep learning is making a profound impact on how computers can discover complex patterns using the self-extracted features from the data (as opposed to a data scientist providing the feature vector to the learning algorithm). Cognitive computing—first popularized by IBM Watson and its success against the best human players in the game show *Jeopardy!*—makes it possible to deal with a new class of problems, the type

of problems that are thought to be solvable only by human ingenuity and creativity, ones that are characterized by ambiguity and uncertainty. This chapter covers the concepts, methods, and application of these two cutting-edge AI technology trends.

6.1 OPENING VIGNETTE: Fighting Fraud with Deep Learning and Artificial Intelligence

THE BUSINESS PROBLEM

Danske Bank is a Nordic universal bank with strong local roots and bridges to the rest of the world. Founded in October 1871, Danske Bank has helped people and businesses in the Nordics realize their ambitions for over 145 years. Its headquarters is in Denmark, with core markets in Denmark, Finland, Norway, and Sweden.

Mitigating fraud is a top priority for banks. According to the Association of Certified Fraud Examiners, businesses lose more than $3.5 trillion each year to fraud. The problem is pervasive across the financial industry and is becoming more prevalent and sophisticated each month. As customers conduct more banking online across a wider variety of channels and devices, there are more opportunities for fraud to occur. Adding to the problem, fraudsters are becoming more creative and technologically savvy—they are also using advanced technologies such as machine learning—and new schemes to defraud banks are evolving rapidly.

Old methods for identifying fraud, such as using human-written rules engines, catch only a small percentage of fraud cases and produce a significantly high number of false positives. While false negatives end up costing money to the bank, chasing after a large number of false positives not only costs time and money but also blemishes customer trust and satisfaction. To improve probability predictions and identify a much higher percentage of actual cases of fraud while reducing false alarms, banks need new forms of analytics. This includes using artificial intelligence.

Danske Bank, like other global banks, is seeing a seismic shift in customer interactions. In the past, most customers handled their transactions in a bank branch. Today, almost all interactions take place digitally through a mobile phone, tablet, ATM, or call center. This provides more "surface area" for fraud to occur. The bank needed to modernize its fraud detection defenses. It struggled with a low 40 percent fraud detection rate and was managing up to 1,200 false positives per day—and 99.5 percent of all cases the bank was investigating were not fraud related. That large number of false alarms required a substantial investment of people, time, and money to investigate what turned out to be dead ends. Working with Think Big Analytics, a Teradata company, Danske Bank made a strategic decision to apply innovative analytic techniques, including AI, to better identify instances of fraud while reducing false positives.

THE SOLUTION: DEEP LEARNING ENHANCES FRAUD DETECTION

Danske Bank integrated deep learning with graphics processing unit (GPU) appliances that were also optimized for deep learning. The new software system helps the analytics team to identify potential cases of fraud while intelligently avoiding false positives. Operational decisions are shifted from users to AI systems. However, human intervention is still necessary in some cases. For example, the model can identify anomalies, such as debit card purchases taking place around the world, but analysts are needed to determine whether that is fraud or a bank customer simply made an online purchase that sent a payment to China and then bought an item the next day from a retailer based in London.

Danske Bank's analytic approach employs a "champion/challenger" methodology. With this approach, deep learning systems compare models in real time to determine which one is most effective. Each challenger processes data in real time, learning as it goes which traits are more likely to indicate fraud. If a process dips below a certain threshold, the model is fed more data, such as the geolocation of customers or recent ATM transactions. When a challenger outperforms other challengers, it transforms into a champion, giving the other models a roadmap to successful fraud detection.

THE RESULTS

Danske Bank implemented a modern enterprise analytic solution leveraging AI and deep learning, and it has paid big dividends. The bank was able to:

- Realize a 60 percent reduction in false positives with an expectation to reach as high as 80 percent.
- Increase true positives by 50 percent.
- Focus resources on actual cases of fraud.

The following graph (see Figure 6.1) shows how true and false positive rates improved with advanced analytics (including deep learning). The red dot represents the old rules engine, which caught only about 40 percent of all fraud. Deep learning improved significantly upon machine learning, allowing Danske Bank to better detect fraud with much lower false positives.

Enterprise analytics is rapidly evolving and moving into new learning systems enabled by AI. At the same time, hardware and processors are becoming more powerful and specialized, and algorithms more accessible, including those available through open source. This gives banks the powerful solutions needed to identify and mitigate fraud. As Danske Bank learned, building and deploying an enterprise-grade analytics solution that meets its specific needs and leverages its data sources deliver more value than traditional off-the-shelf tools could have provided. With AI and deep learning, Danske Bank now has the ability to better uncover fraud without being burdened by an unacceptable amount of false positives. The solution also allows the bank's engineers, data scientists, lines of business, and investigative officers from Interpol, local police, and other agencies to collaborate to uncover fraud, including sophisticated fraud rings. With its enhanced capabilities, the enterprise analytic solution is now being used across other business areas of the bank to deliver additional value.

Because these technologies are still evolving, implementing deep learning and AI solutions can be difficult for companies to achieve on their own. They can benefit by partnering with a company that has the proven capabilities to implement technology-enabled solutions that deliver high-value outcomes. As shown in this case, Think Big Analytics, a Teradata company, has the expertise to configure specialized hardware and software frameworks to enable new operational processes. The project entailed integrating open-source solutions, deploying production models, and then applying deep learning

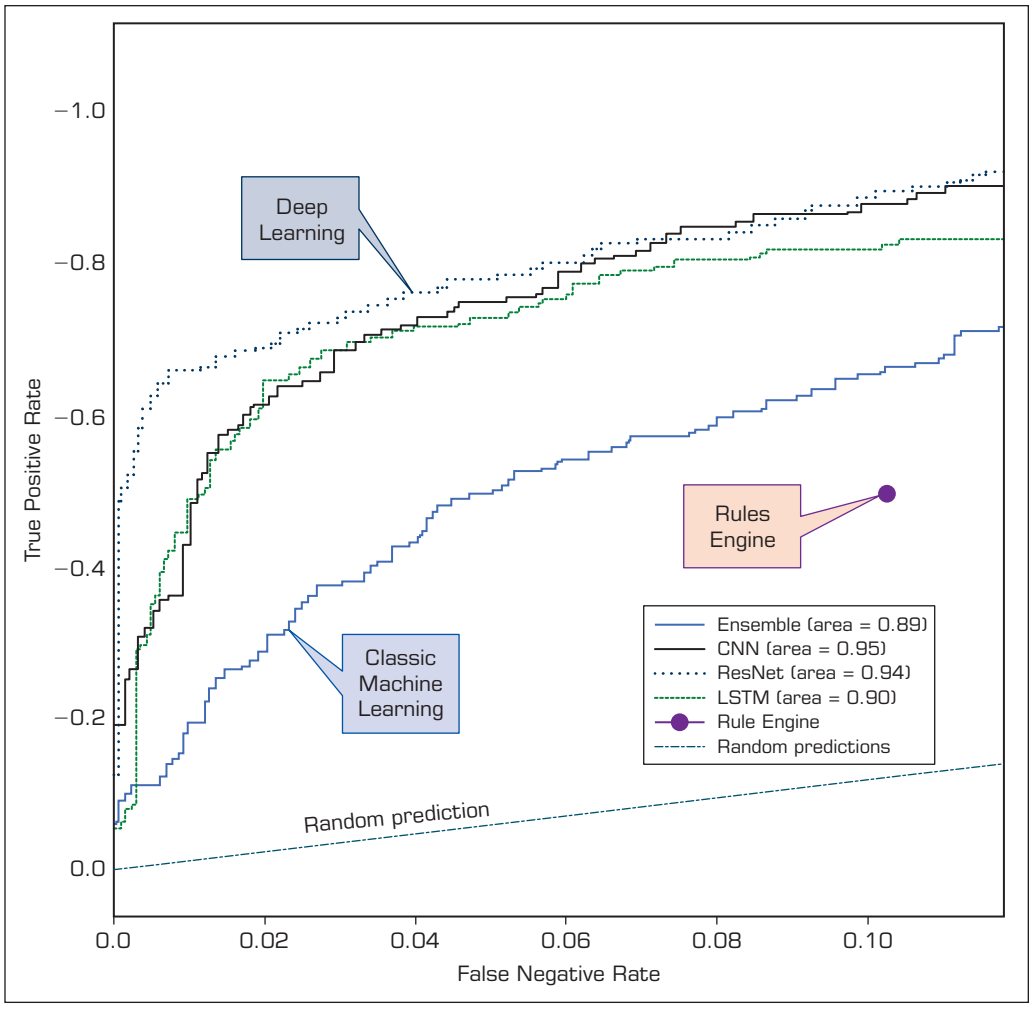


FIGURE 6.1 **Deep Learning Improves Both True Positives and True Negatives.**

analytics to extend and improve the models. A framework was created to manage and track the models in the production system and to make sure the models could be trusted. These models enabled the underlying system to make autonomous decisions in real time that aligned with the bank's procedural, security, and high-availability guidelines. The solution provided new levels of detail, such as time series and sequences of events, to better assist the bank with its fraud investigations. The entire solution was implemented very quickly—from kickoff to live in only five months. Figure 6.2 shows a generalized framework for AI and deep learning–based enterprise-level analytics solutions.

In summary, Danske Bank undertook a multi-step project to productionize machine-learning techniques while developing deep learning models to test those techniques. The integrated models helped identify the growing problem of fraud. For a visual summary, watch the video (**https://www.teradata.com/Resources/Videos/Danske-Bank-Innovating-in-Artificial-Intelligence**) and/or read the blog (**http://blogs.teradata.com/customers/danske-bank-innovating-artificial-intelligence-deep-learning-detect-sophisticated-fraud/**).

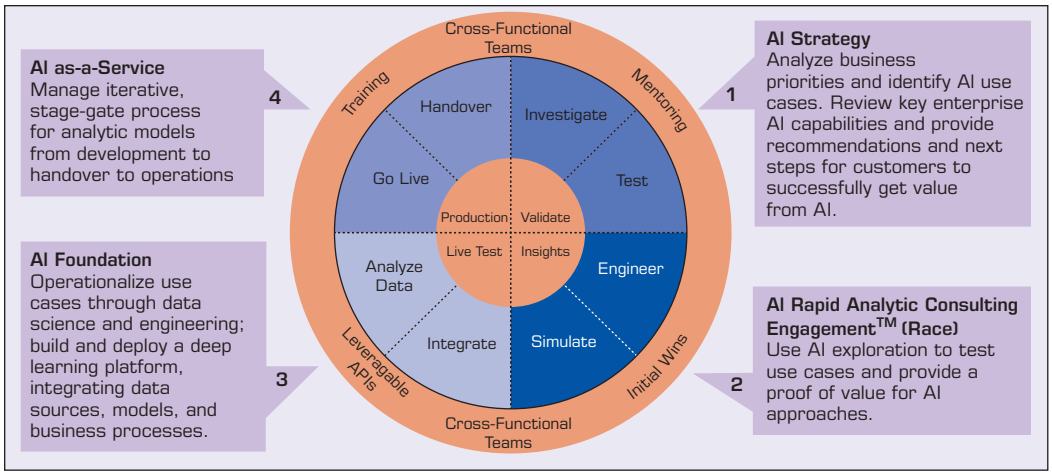


FIGURE 6.2 **A Generalized Framework for AI and Deep Learning–Based Analytics Solutions.**

QUESTIONS FOR THE OPENING VIGNETTE

1. What is fraud in banking?
2. What are the types of fraud that banking firms are facing today?
3. What do you think are the implications of fraud on banks and on their customers?
4. Compare the old and new methods for identifying and mitigating fraud.
5. Why do you think deep learning methods provided better prediction accuracy?
6. Discuss the trade-off between false positive and false negative (type 1 and type 2 errors) within the context of predicting fraudulent activities.

WHAT WE CAN LEARN FROM THIS VIGNETTE

As you will see in this chapter, AI in general and the methods of machine learning in specific are evolving and advancing rapidly. The use of large digitized data sources, both from inside and outside the organization, both structured and unstructured, along with advanced computing systems (software and hardware combinations), has paved the way toward dealing with problems that were thought to be unsolvable just a few years ago. Deep learning and cognitive computing (as the ramifications of the cutting edge in AI systems) are helping enterprises to make accurate and timely decisions by harnessing the rapidly expanding Big Data resources. As shown in this opening vignette, this new generation of AI systems is capable of solving problems much better than their older counterparts. In the domain of fraud detection, traditional methods have always been marginally useful, having higher than desired false positive rates and causing unnecessary investigations and thereby dissatisfaction for their customers. As difficult problems such as fraud detection are, new AI technologies like deep learning are making them solvable with a high level of accuracy and applicability.

Source: Teradata Case Study. "Danske Bank Fights Fraud with Deep Learning and AI." **https://www.teradata.com/Resources/Case-Studies/Danske-Bank-Fight-Fraud-With-Deep-Learning-and-AI** (accessed August 2018). Used with permission.

6.2 INTRODUCTION TO DEEP LEARNING

About a decade ago, conversing with an electronic device (in human language, intelligently) would have been unconceivable, something that could only be seen in SciFi movies. Today, however, thanks to the advances in AI methods and technologies, almost everyone has experienced this unthinkable phenomenon. You probably have already asked Siri or Google Assistant several times to dial a number from your phone address book or to find an address and give you the specific directions while you were driving. Sometimes when you were bored in the afternoon, you may have asked the Google Home or Amazon's Alexa to play some music in your favorite genre on the device or your TV. You might have been surprised at times when you uploaded a group photo of your friends on Facebook and observed its tagging suggestions where the name tags often exactly match your friends' faces in the picture. Translating a manuscript from a foreign language does not require hours of struggling with a dictionary; it is as easy as taking a picture of that manuscript in the Google Translate mobile app and giving it a fraction of a second. These are only a few of the many, ever-increasing applications of deep learning that have promised to make life easier for people.

Deep learning, as the newest and perhaps at this moment the most popular member of the AI and machine-learning family, has a goal similar to those of the other machine-learning methods that came before it: mimic the thought process of humans—using mathematical algorithms to learn from data pretty much the same way that humans learn. So, what is really different (and advanced) in deep learning? Here is the most commonly pronounced differentiating characteristic of deep learning over traditional machine learning. The performance of traditional machine-learning algorithms such as decision trees, support vector machines, logistic regression, and neural networks relies heavily on the representation of the data. That is, only if we (analytics professionals or data scientists) provide those traditional machine-learning algorithms with relevant and sufficient pieces of information (a.k.a. features) in proper format are they able to "learn" the patterns and thereby perform their prediction (classification or estimation), clustering, or association tasks with an acceptable level of accuracy. In other words, these algorithms need humans to manually identify and derive features that are theoretically and/or logically relevant to the objectives of the problem on hand and feed these features into the algorithm in a proper format. For example, in order to use a decision tree to predict whether a given customer will return (or churn), the marketing manager needs to provide the algorithm with information such as the customer's socioeconomic characteristics—income, occupation, educational level, and so on (along with demographic and historical interactions/transactions with the company). But the algorithm itself is not able to define such socioeconomic characteristics and extract such features, for instance, from survey forms completed by the customer or obtained from social media.

While such a structured, human-mediated machine-learning approach has been working fine for rather abstract and formal tasks, it is extremely challenging to have the approach work for some informal, yet seemingly easy (to humans), tasks such as face identification or speech recognition since such tasks require a great deal of knowledge about the world (Goodfellow et al., 2016). It is not straightforward, for instance, to train a machine-learning algorithm to accurately recognize the real meaning of a sentence spoken by a person just by manually providing it with a number of grammatical or semantic features. Accomplishing such a task requires a "deep" knowledge about the world that is not easy to formalize and explicitly present. What deep learning has added to the classic machine-learning methods is in fact the ability to automatically acquire the knowledge required to accomplish such informal tasks and consequently extract some advanced features that contribute to the superior system performance.

To develop an intimate understanding of deep learning, one should learn where it fits in the big picture of all other AI family of methods. A simple hierarchical relationship diagram,

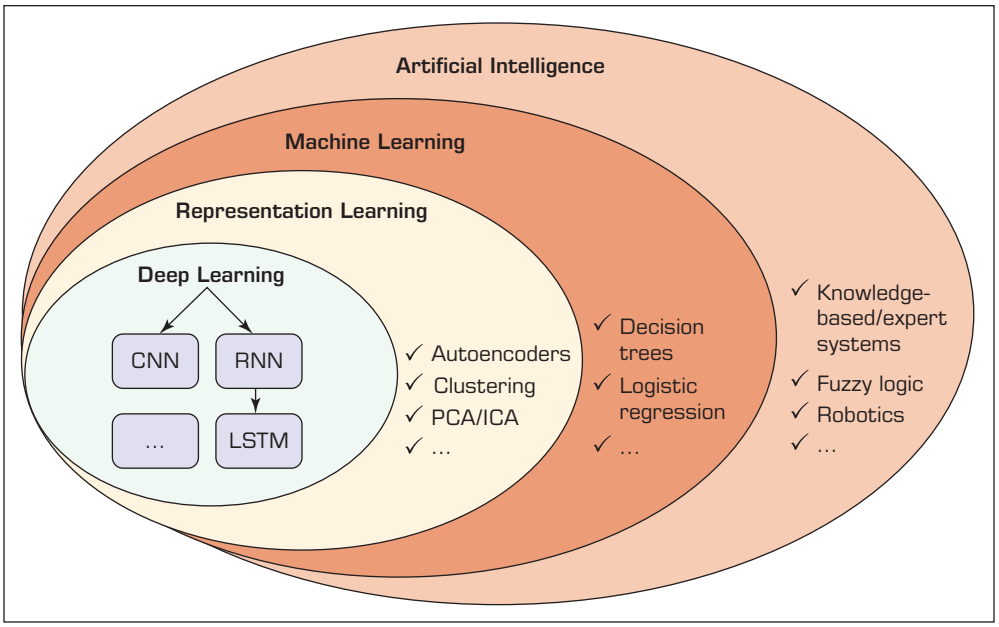


FIGURE 6.3 A Venn Diagram Showing the Placement of Deep Learning within the Overarching AI-Based Learning Methods.

or a taxonomy-like representation, may in fact provide such a holistic understanding. In an attempt to do this, Goodfellow and his colleagues (2016) categorized deep learning as part of the representation learning family of methods. Representation learning techniques entail one type of machine learning (which is also a part of AI) in which the emphasis is on learning and discovering features by the system in addition to discovering the mapping from those features to the output/target. Figure 6.3 uses a Venn diagram to illustrate the placement of deep learning within the overarching family of AI-based learning methods.

Figure 6.4 highlights the differences in the steps/tasks that need to be performed when building a typical deep learning model versus the steps/tasks performed when building models with classic machine-learning algorithms. As shown in the top two workflows, knowledge-based systems and classic machine-learning methods require data scientists to manually create the features (i.e., the representation) to achieve the desired output. The bottommost workflows show that deep learning enables the computer to derive some complex features from simple concepts that would be very effort intensive (or perhaps impossible in some problem situations) to be discovered by humans manually, and then it maps those advanced features to the desired output.

From a methodological viewpoint, although deep learning is generally believed to be a new area in machine learning, its initial idea goes back to the late 1980s, just a few decades after the emergence of artificial neural networks when LeCun and colleagues (1989) published an article about applying backpropagation networks for recognizing handwritten ZIP codes. In fact, as it is being practiced today, deep learning seems to be nothing but an extension of neural networks with the idea that deep learning is able to deal with more complicated tasks with a higher level of sophistication by employing many layers of connected neurons along with much larger data sets to automatically characterized variables and solve the problems but only at the expense of a great deal of computational effort. This very high computational requirement and the need for very large data sets were the two main reasons why the initial idea had to wait more than two decades until some advanced computational and technological infrastructure emerged for deep

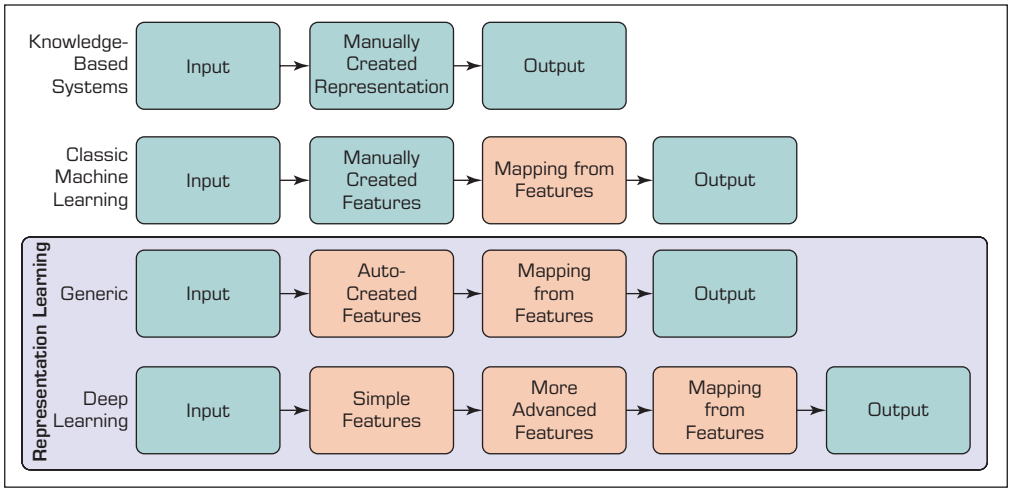


FIGURE 6.4 **Illustration of the Key Differences between Classic Machine-Learning Methods and Representation Learning/Deep Learning (shaded boxes indicate components that are able to learn directly from data).**

learning's practical realization. Although the scale of neural networks has dramatically increased in the past decade by the advancement of related technologies, it is still estimated that having artificial deep neural networks with the comparable number of neurons and level of complexity existing in the human brain will take several more decades.

In addition to the computer infrastructures, as mentioned, the availability of large and feature-rich digitized data sets was another key reason for the development of successful deep learning applications in recent years. Obtaining good performance from a deep learning algorithm used to be a very difficult task that required extensive skills and experience/understanding to design task-specific networks, and therefore, not many were able to develop deep learning for practical and/or research purposes. Large training data sets, however, have greatly compensated for the lack of intimate knowledge and reduced the level of skill needed for implementing deep neural networks. Nevertheless, although the size of available data sets has exponentially increased in recent years, a great challenge, especially for supervised learning of deep networks, is now the labeling of the cases in these huge data sets. As a result, a great deal of research is ongoing, focusing on how we can take advantage of large quantities of unlabeled data for semisupervised or unsupervised learning or how we can develop methods to label examples in bulk in a reasonable time.

The following section of this chapter provides a general introduction to neural networks from where deep learning has originated. Following the overview of these "shallow" neural networks, the chapter introduces different types of deep learning architectures and how they work, some common applications of these deep learning architectures, and some popular computer frameworks to use in implementing deep learning in practice. Since, as mentioned, the basics of deep learning are the same as those of artificial neural networks, in the following section, we provide a brief coverage of the neural network architecture (namely, multilayered perceptron [MLP]-type neural networks, which was omitted in the neural network section in Chapter 5 because it was to be covered here) to focus on their mathematical principles and then explain how the various types of deep learning architectures/approaches were derived from these foundations. Application Case 6.1 provides an interesting example of what deep learning and advanced analytics techniques can achieve in the field of football.

Application Case 6.1 Finding the Next Football Star with Artificial Intelligence

Football. Soccer. The beautiful game. Whatever you call it, the world's most popular sport is being transformed by a Dutch start-up bringing AI to the pitch. SciSports, founded in 2012 by two self-proclaimed football addicts and data geeks, is innovating on the edge of what is possible. The sports analytics company uses streaming data and applies machine learning, deep learning, and AI to capture and analyze these data, making way for innovations in everything from player recruitment to virtual reality for fans.

Player Selection Goes High Tech

In the era of eight-figure contracts, player recruitment is a high-stakes game. The best teams are not those with the best players but the best combination of players. Scouts and coaches have used observation, rudimentary data, and intuition for decades, but savvy clubs now are using advanced analytics to identify rising stars and undervalued players. "The SciSkill Index evaluates every professional football player in the world in one universal index," says SciSports founder and CEO Giels Brouwer. The company uses machine-learning algorithms to calculate the quality, talent, and value of more than 200,000 players. This helps clubs find talent, look for players who fit a certain profile, and analyze their opponents.

Every week, more than 1,500 matches in 210 leagues are analyzed by the SciSkill technology. Armed with this insight, SciSports partners with elite football clubs across Europe and other continents to help them sign the right players. This has led to several unexpected—and in some cases lucrative—player acquisitions. For example, a second-division Dutch player did not want to renew his contract, so he went out as a free agent. A new club reviewed the SciSkill index and found his data intriguing. That club was not too sure at first because it thought he looked clumsy in scouting—but the data told the true story. The club signed him as the third striker, and he quickly moved into a starting role and became its top goal scorer. His rights were sold at a large premium within two years, and now he is one of the top goal scorers in Dutch professional football.

Real-Time 3D Game Analysis

Traditional football data companies generate data only on players who have the ball, leaving everything else

(*Continued*)

Application Case 6.1 (Continued)

undocumented. This provides an incomplete picture of player quality. Seeing an opportunity to capture the immense amount of data regarding what happens away from the ball, SciSports developed a camera system called BallJames.

BallJames is a real-time tracking technology that automatically generates 3D data from video. Fourteen cameras placed around a stadium record every movement on the field. BallJames then generates data such as the precision, direction, and speed of the passing, sprinting strength, and jumping strength. "This forms a complete picture of the game," says Brouwer. "The data can be used in lots of cool ways, from allowing fans to experience the game from any angle using virtual reality, to sports betting and fantasy sports." He added that the data can even help coaches on the bench. "When they want to know if a player is getting tired, they can substitute players based on analytics."

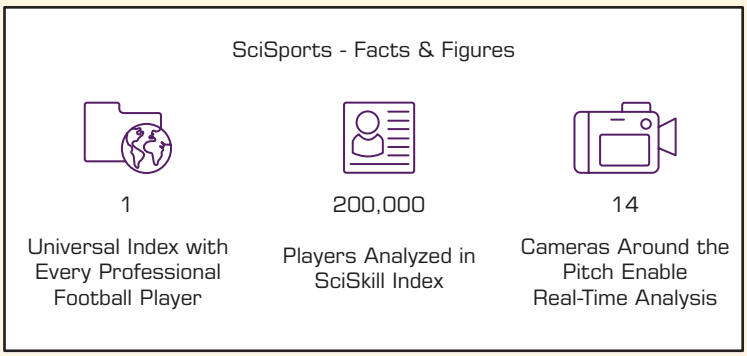
Machine Learning and Deep Learning

SciSports models on-field movements using machine-learning algorithms, which by nature improve on performing a task as the player gains more experience. On the pitch, BallJames works by automatically assigning a value to each action, such as a corner kick. Over time, these values change based on their success rate. A goal, for example, has a high value, but a contributing action—which may have previously had a low value—can become more valuable as the platform masters the game. Wouter Roosenburg, SciSports chief technology officer, says AI and machine learning will play an important role in the future of SciSports and football analytics in general. "Existing mathematical models model existing knowledge and insights in football, while artificial intelligence and machine learning will make it possible to discover new connections that people wouldn't make themselves."

To accurately compile 3D images, BallJames must distinguish between players, referees, and the ball. SAS Event Stream Processing enables real-time image recognition using deep learning models. "By combining our deep learning models into SAS®Viya®, we can train our models in-memory in the cloud, on our cameras or wherever our resources are," says Roosenburg. The ability to deploy deep learning models in memory onto cameras and then do the inferencing in real time is cutting-edge science. "Having one uniform platform to manage the entire 3-D production chain is invaluable," says Roosenburg. "Without SAS Viya, this project would not be possible."

Adding Oomph to Open Source

Previously SciSports exclusively used open source to build models. It now benefits from an end-to-end platform that allows analytical teams to work in their language of choice and share a single, managed analytical asset inventory across the organization. According to Brouwer, this enables the firm to attract employees with different open-source skills yet still manage the production chain using one platform. "My CTO tells me he loves that our data scientists can do all the research in open source and he doesn't have to worry about the production of the models," says Brouwer. "What takes 100 lines of code in Python only takes five in SAS. This speeds our time to market, which is crucial in sports analytics."



Since its inception, SciSports has quickly become one of the world's fastest-growing sports analytics companies. Brouwer says the versatility of the SAS Platform has also been a major factor. "With SAS, we've got the ability to scale processing power up or down as needed, put models into production in real time, develop everything in one platform and integrate with open source. Our ambition is to bring real-time data analytics to billions of soccer fans all over the world. By partnering with SAS, we can make that happen."

Questions for Case 6.1

1. What does SciSports do? Look at its Web site for more information.
2. How can advanced analytics help football teams?
3. What is the role of deep learning in solutions provided by SciSports?

Sources: SAS Customer Stories. "Finding the Next Football Star with Artificial Intelligence." **www.sas.com/en_us/customers/scisports.html** (accessed August 2018).Copyright (c) 2018 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Used with permission.

SECTION 6.2 REVIEW QUESTIONS

1. What is deep learning? What can deep learning do?
2. Compared to traditional machine learning, what is the most prominent difference of deep learning?
3. List and briefly explain different learning methods in AI.
4. What is representation learning, and how does it relate to deep learning?

6.3 BASICS OF "SHALLOW" NEURAL NETWORKS

Artificial neural networks are essentially simplified abstractions of the human brain and its complex biological networks of neurons. The human brain has a set of billions of interconnected neurons that facilitate our thinking, learning, and understanding of the world around us. Theoretically speaking, learning is nothing but the establishment and adaptation of new or existing interneuron connections. In the artificial neural networks, however, neurons are processing units (also called **processing elements [PEs]**) that perform a set of predefined mathematical operations on the numerical values coming from the input variables or from the other neuron outputs to create and push out its own outputs. Figure 6.5 shows a schematic representation of a single-input and single-output neuron (more accurately, the processing element in artificial neural networks).

In this figure, p represents a numerical input. Each input goes into the neuron with an *adjustable* weight w and a bias term b. A multiplication *weight function* applies the weight to the input, and a *net input function* shown by Σ adds the bias term to the weighted input z. The output of the net input function (n, known as the *net input*) then goes through another function called the *transfer* (a.k.a. activation) function (shown by f) for conversion and the production of the actual output a. In other words:

$$a = f(wp + b)$$

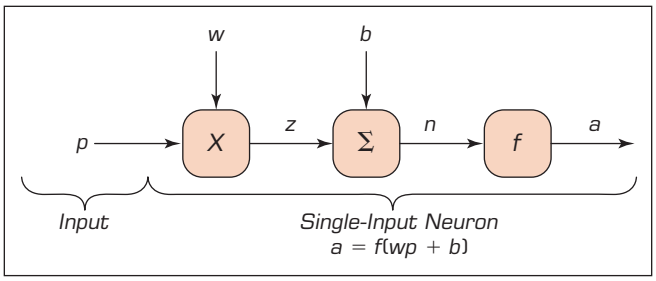


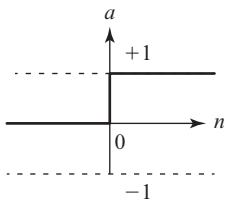
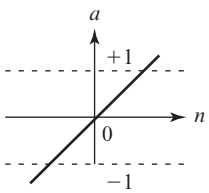
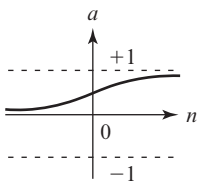
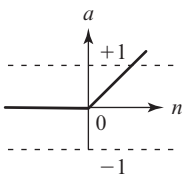
FIGURE 6.5 General Single-Input Artificial Neuron Representation.

A numerical example: if $w = 2$, $p = 3$, and $b = -1$, then $a = f(2 * 3 - 1) = f(5)$.

Various types of transfer functions are commonly used in the design of neural networks. Table 6.1 shows some of the most common transfer functions and their corresponding operations. Note that in practice, selection of proper transfer functions for a network requires a broad knowledge of neural networks—characteristics of the data as well as the specific purpose for which the network is created.

Just to provide an illustration, if in the previous example we had a hard limit transfer function, the actual output a would be $a = hardlim(5) = 1$. There are some guidelines for choosing the appropriate transfer function for each set of neurons in a network. These guidelines are especially robust for the neurons located at the output layer of the network. For example, if the nature of the output for a model is binary, we are advised to use *Sigmoid* transfer functions at the output layer so that it produces an output between 0 and 1, which represents the conditional probability of $y = 1$ given x *or* $P(y = 1|x)$. Many neural network textbooks provide and elaborate on those guidelines at different layers in a neural network with some consistency and much disagreement, suggesting that the best practices should (and usually does) come from experience.

TABLE 6.1 Common Transfer (Activation) Functions in Neural Networks

Transfer Function	Form	Operation
Hard limit	 $a = hardlim\,(n)$	$a = +1$ *if* $n > 0$ $a = 0$ *if* $n < 0$
Linear	 $a = purelin\,(n)$	$a = n$
Log-Sigmoid	 $a = logsig\,(n)$	$a = \frac{1}{1 + e^{-n}}$
Positive linear (a.k.a. rectified linear or ReLU)	 $a = poslin(n)$	$a = n$ *if* $n > 0$ $a = 0$ *if* $n < 0$

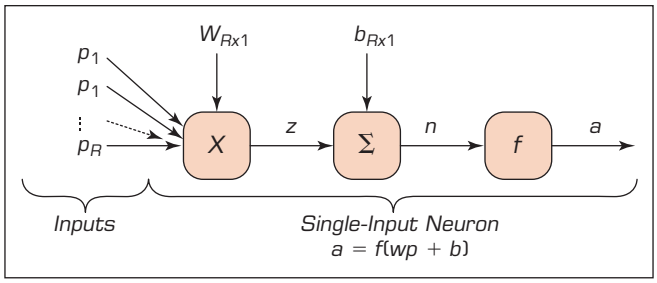


FIGURE 6.6 **Typical Multiple-Input Neuron with R Individual Inputs.**

Typically, a neuron has more than a single input. In that case, each individual input p_i can be shown as an element of the input vector **p**. Each of the individual input values would have its own adjustable weight w_i of the weight vector **W**. Figure 6.6 shows a multiple-input neuron with R individual inputs.

For this neuron, the net input n can be expressed as:

$$n = w_{1,1}p_1 + w_{1,2}p_2 + w_{1,3}p_3 + \ldots + w_{1,R}p_R + b$$

Considering the input vector **p** as a $R \times 1$ vector and the weight vector **W** as a $1 \times$ R vector, then n can be written in matrix form as:

$$n = \mathbf{Wp} + b$$

where **Wp** is a scalar (i.e., 1×1 vector).

Moreover, each neural network is typically composed of multiple neurons connected to each other and structured in consecutive *layers* so that the outputs of a layer work as the inputs to the next layer. Figure 6.7 shows a typical neural network with four neurons

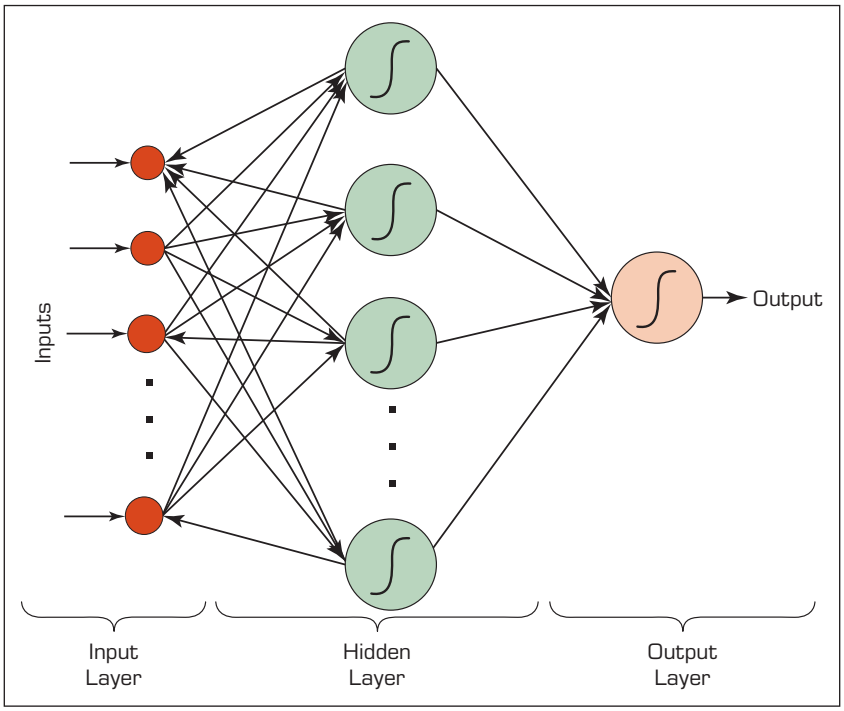


FIGURE 6.7 **Typical Neural Network with Three Layers and Eight Neurons.**

at the input (i.e., first) layer, four neurons at the hidden (i.e., middle) layer, and a single neuron at the output (i.e., last) layer. Each of the neurons has its own weight, weighting function, bias, and transfer function and processes its own input(s) as described.

While the inputs, weighting functions, and transfer functions in a given network are fixed, the values of the weights and biases are adjustable. The process of adjusting weights and biases in a neural network is what is commonly called *training*. In fact, in practice, a neural network cannot be used effectively for a prediction problem unless it is well trained by a sufficient number of examples with known *actual outputs* (a.k.a. *targets*). The goal of the training process is to adjust network weights and biases such that the network output for each set of inputs (i.e., each sample) is adequately close to its corresponding target value.

Application Case 6.2 provides a case where computer gaming companies are using advanced analytics to better understand and engage with their customers.

Application Case 6.2

Gaming Companies Use Data Analytics to Score Points with Players

Video gamers are a special breed. Sure, they spend a lot of time playing games, but they're also building social networks. Like sports athletes, video game players thrive on competition. They play against other gamers online. Those who earn first place, or even second or third place, have bragging rights. And like athletes who invest a lot of time training, video gamers take pride in the number of hours they spend playing. Furthermore, as games increase in complexity, gamers take pride in developing unique skills to best their compatriots.

A New Level of Gaming

Video gaming has evolved from the days of PAC-MAN and arcades. The widespread availability of the Internet has fueled the popularity of video games by bringing them into people's homes via a wide range of electronics such as the personal computer and mobile devices. The world of computer games is now a powerful and profitable business.

According to NewZoo's *Global Games Market Report* from April 2017, the global games market in 2017 saw:

- $109 billion in revenues.
- 7.8 percent increase from the previous year.
- 2.2 billion gamers globally.
- 42 percent of the market being mobile.

Video game companies can tap into this environment and learn valuable information about their customers, especially their behaviors and the underlying motivations. These customer data enable companies to improve the gaming experience and better engage players.

Traditionally, the gaming industry appealed to its customers—the gamers—by offering striking graphics and captivating visualizations. As technology advanced, the graphics became more vivid with hi-def renditions. Companies have continued to use technology in highly creative ways to develop games that attract customers and capture their interests, which results in more time spent playing and higher affinity levels. What video game companies have not done as well is to fully utilize technology to understand the contextual factors that drive sustained brand engagement.

Know the Players

In today's gaming world, creating an exciting product is no longer enough. Games must strongly appeal to the visual and auditory senses in an era when people expect cool graphics and cutting-edge sound effects. Games must also be properly marketed to reach highly targeted player groups. There are also opportunities to monetize gaming characters in the form of commercially available merchandise (e.g., toy store characters) or movie rights. Making a game successful requires programmers, designers,

scenarists, musicians, and marketers to work together and share information. That is where gamer and gaming data come into play.

For example, the size of a gamer's network—the number and types of people a gamer plays with or against—usually correlates with more time spent playing and more money that is spent. The more relationships gamers have, the higher the likelihood they will play more games with more people because they enjoy the experience. Network effects amplify engagement volumes.

These data also help companies better understand the types of games each individual likes to play. These insights enable the company to recommend additional games across other genres that will likely exert a positive impact on player engagement and satisfaction. Companies can also use these data in marketing campaigns to target new gamers or entice existing gamers to upgrade their memberships, for example, to premium levels.

Monetize Player Behaviors

Collaborative filtering (cFilter) is an advanced analytic function that makes automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The cFilter function supposes that if User A has the same opinion as User B on one issue, then User A is more likely to have User B's opinion on a different issue when compared to a random user. This shows that predictions are specific to a gamer based on data from many other gamers.

Filtering systems are often used by online retailers to make product recommendations. The analytics can determine products that a customer will like based on what other shoppers who made similar purchases also bought, liked, or rated highly. There are many examples across other industries such as healthcare, finance, manufacturing, and telecommunication.

The cFilter analytic function offers several benefits to online video game companies:

- ***Marketers can run more effective campaigns.*** Connections between gamers naturally form to create clusters. Marketers can isolate common player characteristics and leverage those insights for campaigns. Conversely, they can isolate players who do not belong to a cluster and determine what unique characteristics contribute to their nonconforming behaviors.
- ***Companies can improve player retention.*** A strong membership in a community of gamers decreases the chances of churn. The greater the incentives for gamers to belong to a group of active participants, the more desire they have to engage in competitions. This increases the "stickiness" of customers and can lead to more game subscriptions.
- ***Data insights lead to improved customer satisfaction.*** Clusters indicate a desire for certain types of games that correspond to distinct gamer interests and behaviors. Companies can create gaming experiences that are unique to each player. Enticing more people to play and play longer enhances gamer satisfaction.

Once they understand why customers want to play games and uncover their relationships with other gamers, companies can create the right incentives for players to keep returning. This ensures a sustained customer base and stable revenue streams.

Boost Loyalty and Revenue

Regardless of the genre, each video game has passionate players who seek each other for competitions. The thrill of a conquest attracts avid engagement. Over time, distinct networks of gamers are formed, with each participant constructing social relationships that often lead to more frequent and intense gaming interactions.

The gaming industry is now utilizing data analytics and visualizations to discern customer behaviors better and uncover player motivations. Looking at customer segments is no longer enough. Companies are now looking at microsegments that go beyond traditional demographics like age or geographic location to understand customer preferences such as favorite games, preferred levels of difficulty, or game genres.

By gaining analytic insights into gamer strategies and behaviors, companies can create unique

(*Continued*)

Application Case 6.2 (Continued)

gaming experiences that are attuned to these behaviors. By engaging players with the games and features they desire, video game companies gain a devoted following, grow profits, and develop new revenue streams through merchandising ventures.

For a visual treat, watch a short video (**https://www.teradata.com/Resources/Videos/Art-of-Analytics-The-Sword**) to see how the companies can use analytics to decipher gamer relationships that drive user behaviors and lead to better games.

Questions for Case 6.2

1. What are the main challenges for gaming companies?
2. How can analytics help gaming companies stay competitive?
3. What types of data can gaming companies obtain and use for analytics?

Source: Teradata Case Study. **https://www.teradata.com/Resources/Case-Studies/Gaming-Companies-Use-Data-Analytics** (accessed August 2018).

Technology Insight 6.1 briefly describes the common components (or elements) of a typical artificial neural network along with their functional relationships.

TECHNOLOGY INSIGHT 6.1 Elements of an Artificial Neural Network

A neural network is composed of processing elements that are organized in different ways to form the network's structure. The basic processing unit in a neural network is the neuron. A number of neurons are then organized to establish a network of neurons. Neurons can be organized in a number of different ways; these various network patterns are referred to as *topologies* or *network architectures* (some of the most common architectures are summarized in Chapter 5). One of the most popular approaches, known as the *feedforward-multilayered perceptron*, allows all neurons to link the output in one layer to the input of the next layer, but it does not allow any feedback linkage (Haykin, 2009).

Processing Element (PE)

The PE of an ANN is an artificial neuron. Each neuron receives inputs, processes them, and delivers a single output as shown in Figure 6.5. The input can be raw input data or the output of other processing elements. The output can be the final result (e.g., 1 means yes, 0 means no), or it can be input to other neurons.

Network Structure

Each ANN is composed of a collection of neurons that are grouped into layers. A typical structure is shown in Figure 6.8. Note the three layers: input, intermediate (called *the hidden layer)*, and output. A **hidden layer** is a layer of neurons that takes input from the previous layer and converts those inputs into outputs for further processing. Several hidden layers can be placed between the input and output layers, although it is common to use only one hidden layer. In that case, the hidden layer simply converts inputs into a nonlinear combination and passes the transformed inputs to the output layer. The most common interpretation of the hidden layer is as a feature-extraction mechanism; that is, the hidden layer converts the original inputs in the problem into a higher-level combination of such inputs.

In ANN, when information is processed, many of the processing elements perform their computations at the same time. This parallel processing resembles the way the human brain works, and it differs from the serial processing of conventional computing.

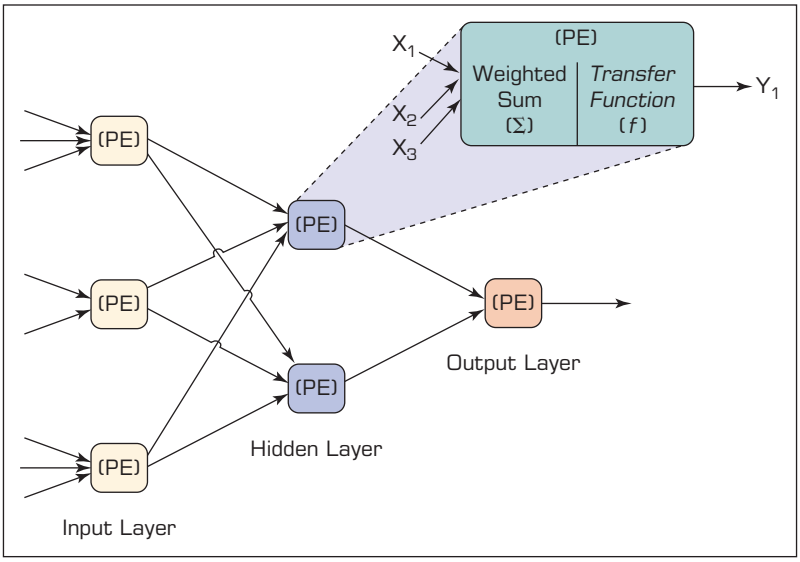


FIGURE 6.8 Neural Network with One Hidden Layer. PE: processing element (an artificial representation of a biological neuron); X_i: inputs to a PE; y: output generated by a PE; $\sum$: summation function; and f: activation/transfer function.

Input

Each input corresponds to a single attribute. For example, if the problem is to decide on approval or disapproval of a loan, attributes could include the applicant's income level, age, and home ownership status. The numeric value, or the numeric representation of non-numeric value, of an attribute is the input to the network. Several types of data, such as text, picture, and voice, can be used as inputs. Preprocessing may be needed to convert the data into meaningful inputs from symbolic/non-numeric data or to numeric/scale data.

Outputs

The output of a network contains the solution to a problem. For example, in the case of a loan application, the output can be "yes" or "no." The ANN assigns numeric values to the output, which may then need to be converted into categorical output using a threshold value so that the results would be 1 for "yes" and 0 for "no."

Connection Weights

Connection weights are the key elements of an ANN. They express the relative strength (or mathematical value) of the input data or the many connections that transfer data from layer to layer. In other words, weights express the relative importance of each input to a processing element and, ultimately, to the output. Weights are crucial in that they store learned patterns of information. It is through repeated adjustments of weights that a network learns.

Summation Function

The summation function computes the weighted sums of all input elements entering each processing element. A summation function multiplies each input value by its weight and totals the values for a weighted sum. The formula for n inputs (represented with X) in one processing element is shown in Figure 6.9a, and for several processing elements, the summation function formulas are shown in Figure 6.9b.

Transfer Function

The summation function computes the internal stimulation, or activation level, of the neuron. Based on this level, the neuron may or may not produce an output. The relationship between the

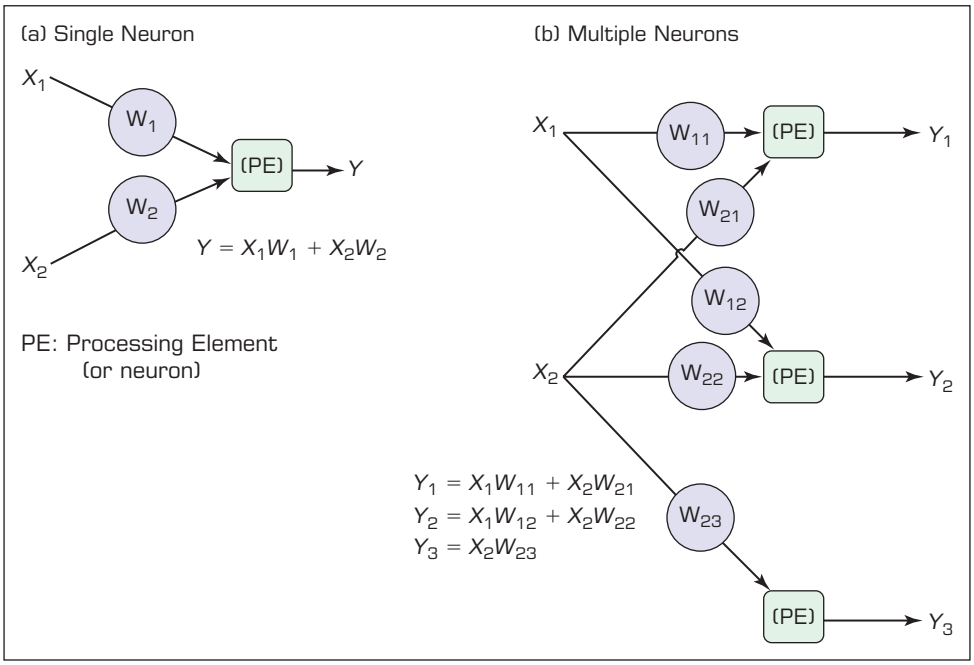


FIGURE 6.9 **Summation Function for (a) a Single Neuron/PE and (b) Several Neurons/PEs.**

internal activation level and the output can be linear or nonlinear. The relationship is expressed by one of several types of transformation (transfer) functions (see Table 6.1 for a list of commonly used activation functions). Selection of the specific activation function affects the network's operation. Figure 6.10 shows the calculation for a simple sigmoid-type activation function example.

The transformation modifies the output levels to fit within a reasonable range of values (typically between 0 and 1). This transformation is performed before the output reaches the next level. Without such a transformation, the value of the output becomes very large, especially when there are several layers of neurons. Sometimes a threshold value is used instead of a transformation function. A threshold value is a hurdle value for the output of a neuron to trigger the next level of neurons. If an output value is smaller than the threshold value, it will not be passed to the next level of neurons. For example, any value of 0.5 or less becomes 0, and any value above 0.5 becomes 1. A transformation can occur at the output of each processing element, or it can be performed only at the final output nodes.

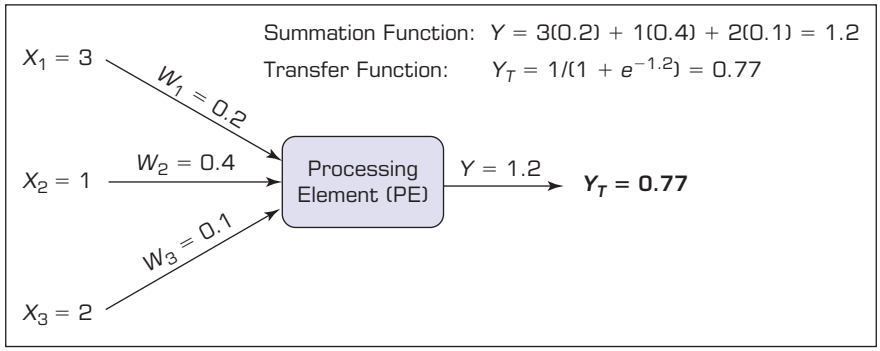


FIGURE 6.10 **Example of ANN Transfer Function.**

Application Case 6.3 provides an interesting use case where advanced analytics and deep learning are being used to prevent the extinction of rare animals.

Application Case 6.3 Artificial Intelligence Helps Protect Animals from Extinction

"There are some people who want to kill animals like the lions and cheetahs. We would like to teach them, there are not many left," says WildTrack officials. The more we can study their behavior, the more we can help to protect them—and sustain the earth's biodiversity that supports us all. Their tracks tell a collective story that holds incredible value in conservation. Where are they going? How many are left? There is much to be learned by monitoring footprints of endangered species like the cheetah.

WildTrack, a nonprofit organization, was founded in 2004 by Zoe Jewell and Sky Alibhai, a veterinarian and a wildlife biologist, respectively, who had been working for many years in Africa monitoring black and white rhinos. While in Zimbabwe, in the early 1990s, they collected and presented data to show that invasive monitoring techniques used for black rhinos were negatively impacting female fertility and began to develop a footprint identification technique. Interest from researchers around the world who needed a cost-effective and noninvasive approach to wildlife monitoring sparked WildTrack.

Artificial intelligence may help people recreate some of the skills used by indigenous trackers. WildTrack researchers are exploring the value AI can bring to conservation. They think that AI solutions are designed to enhance human efforts—not replace them. With deep learning, given enough data, a computer can be trained to perform human-like tasks such as identifying footprint images and recognizing patterns in a similar way to indigenous trackers—but with the added ability to apply these concepts at a much larger scale and more rapid pace. Analytics really underpins the whole thing, potentially giving insights into species populations that WildTrack never had before.

The WildTrack footprint identification technique is a tool for noninvasive monitoring of endangered species through digital images of footprints. Measurements from these images are analyzed by customized mathematical models that help to identify the species, individual, sex, and age class. AI could add the ability to adapt through progressive learning algorithms and tell an even more complete story.

Obtaining crowdsourcing data is the next important step toward redefining what conservation looks like in the future. Ordinary people would not necessarily be able to dart a rhino, but they can take an image of a footprint. WildTrack has data coming in from everywhere—too much to manage traditionally. That's really where AI comes in. It can automate repetitive learning through data, performing frequent, high-volume, computerized tasks reliably and without fatigue.

"Our challenge is how to harness artificial intelligence to create an environment where there's room for us, and all species in this world," says Alibhai.

Questions for Case 6.3

1. What is WildTrack and what does it do?
2. How can advanced analytics help WildTrack?
3. What are the roles that deep learning plays in this application case?

Source: SAS Customer Story. "Can Artificial Intelligence Help Protect These Animals from Extinction? The Answer May Lie in Their Footprints." **https://www.sas.com/en_us/explore/analytics-in-action/impact/WildTrack.html** (accessed August 2018); **WildTrack.org**.

SECTION 6.3 REVIEW QUESTIONS

1. How does a single artificial neuron (i.e., PE) work?
2. List and briefly describe the most commonly used ANN activation functions.
3. What is MLP, and how does it work?
4. Explain the function of weights in ANN.
5. Describe the summation and activation functions in MLP-type ANN architecture.

6.4 PROCESS OF DEVELOPING NEURAL NETWORK–BASED SYSTEMS

Although the development process of ANN is similar to the structured design methodologies of traditional computer-based information systems, some phases are unique or have some unique aspects. In the process described here, we assume that the preliminary steps of system development, such as determining information requirements, conducting a feasibility analysis, and gaining a champion in top management for the project, have been completed successfully. Such steps are generic to any information system.

As shown in Figure 6.11, the development process for an ANN application includes nine steps. In step 1, the data to be used for training and testing the network

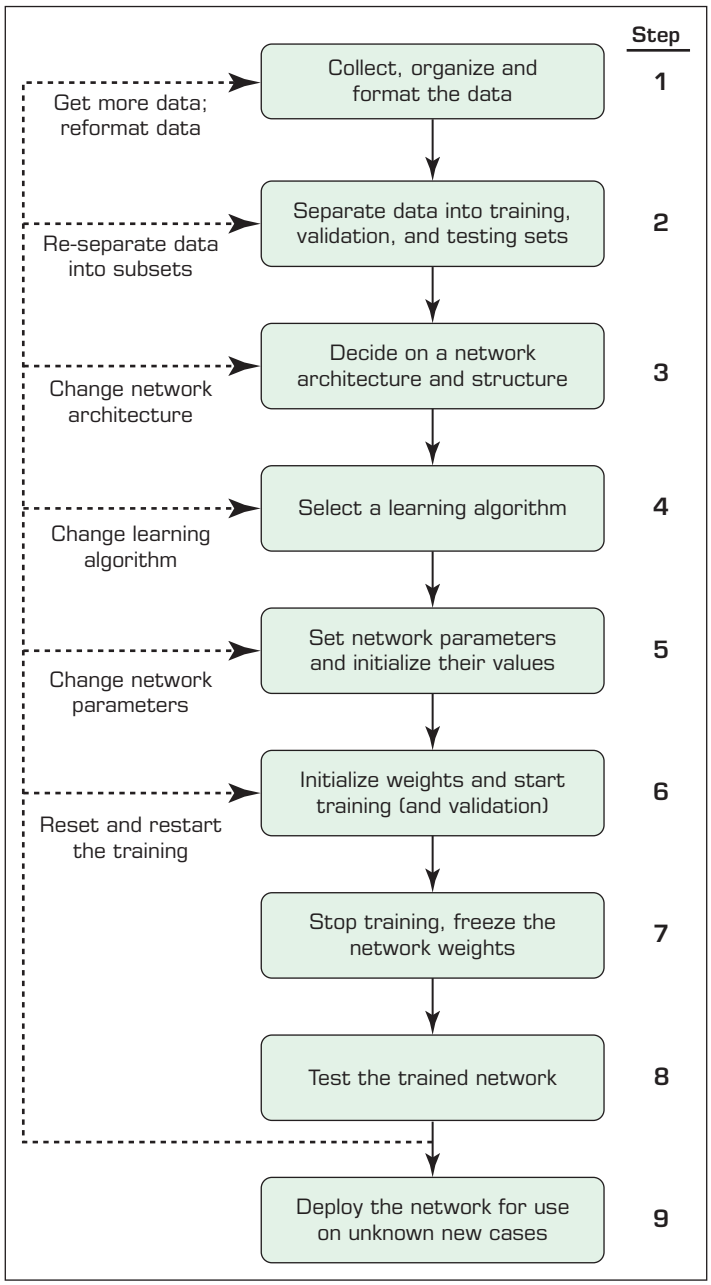


FIGURE 6.11 Development Process of an ANN Model.

are collected. Important considerations are that the particular problem is amenable to a neural network solution and that adequate data exist and can be obtained. In step 2, training data must be identified, and a plan must be made for testing the performance of the network.

In steps 3 and 4, a network architecture and a learning method are selected. The availability of a particular development tool or the capabilities of the development personnel may determine the type of neural network to be constructed. Also, certain problem types have demonstrated high success rates with certain configurations (e.g., multilayer feedforward neural networks for bankruptcy prediction [Altman (1968), Wilson and Sharda (1994), and Olson, Delen, and Meng (2012)]). Important considerations are the exact number of neurons and the number of layers. Some packages use genetic algorithms to select the network design.

There are several parameters for tuning the network to the desired learning performance level. Part of the process in step 5 is the initialization of the network weights and parameters followed by the modification of the parameters as training performance feedback is received. Often, the initial values are important in determining the efficiency and length of training. Some methods change the parameters during training to enhance performance.

Step 6 transforms the application data into the type and format required by the neural network. This may require writing software to preprocess the data or performing these operations directly in an ANN package. Data storage and manipulation techniques and processes must be designed for conveniently and efficiently retraining the neural network when needed. The application data representation and ordering often influence the efficiency and possibly the accuracy of the results.

In steps 7 and 8, training and testing are conducted iteratively by presenting input and desired or known output data to the network. The network computes the outputs and adjusts the weights until the computed outputs are within an acceptable tolerance of the known outputs for the input cases. The desired outputs and their relationships to input data are derived from historical data (i.e., a portion of the data collected in step 1).

In step 9, a stable set of weights is obtained. Then the network can reproduce the desired outputs given inputs such as those in the training set. The network is ready for use as a stand-alone system or as part of another software system where new input data will be presented to it and its output will be a recommended decision.

Learning Process in ANN

In **supervised learning**, the learning process is inductive; that is, connection weights are derived from existing cases. The usual process of learning involves three tasks (see Figure 6.12):

1. Compute temporary outputs.
2. Compare outputs with desired targets.
3. Adjust the weights and repeat the process.

Like any other supervised machine-learning technique, neural network training is usually done by defining a **performance function** (F) (a.k.a. *cost function* or *loss function*) and optimizing (minimizing) that function by changing model parameters. Usually, the performance function is nothing but a measure of error (i.e., the difference between the actual input and the target) across all inputs of a network. There are several types of error measures (e.g., sum square errors, mean square errors, cross entropy, or even custom measures) all of which are designed to capture the difference between the network outputs and the actual outputs.

The training process begins by calculating outputs for a given set of inputs using some random weights and biases. Once the network outputs are on hand, the performance

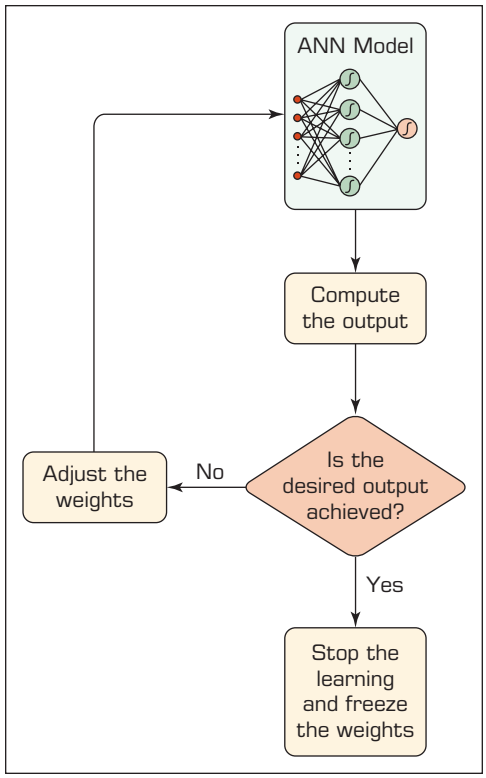


FIGURE 6.12 **Supervised Learning Process of an ANN.**

function can be computed. The difference between the actual output (Y or Y_T) and the desired output (Z) for a given set of inputs is an error called *delta* (in calculus, the Greek symbol delta, Δ, means "difference").

The objective is to minimize delta (i.e., reduce it to 0 if possible), which is done by adjusting the network's weights. The key is to change the weights in the proper direction, making changes that reduce delta (i.e., error). Different ANNs compute delta in different ways, depending on the learning algorithm being used. Hundreds of learning algorithms are available for various situations and configurations of ANN.

Backpropagation for ANN Training

The optimization of performance (i.e., minimization of the error or delta) in the neural network is usually done by an algorithm called **stochastic gradient descent (SGD)**, which is an iterative gradient-based optimizer used for finding the minimum (i.e., the lowest point) in performance functions, as in the case of neural networks. The idea behind the SGD algorithm is that the derivative of the performance function with respect to each current weight or bias indicates the amount of change in the error measure by each unit of change in that weight or bias element. These derivatives are referred to as *network gradients*. Calculation of network gradients in the neural networks requires application of an algorithm called **backpropagation**, which is the most popular neural network learning algorithm, that applies *the chain rule of calculus* to compute the derivatives of functions formed by composing other functions whose derivatives are known [more on the mathematical details of this algorithm can be found in Rumelhart, Hinton, and Williams (1986)].

Backpropagation (short for *back-error propagation*) is the most widely used supervised learning algorithm in neural computing (Principe, Euliano, and Lefebvre, 2000). By using the SGD mentioned previously, the implementation of backpropagation algorithms is relatively straightforward. A neural network with backpropagation learning includes one or more hidden layers. This type of network is considered feedforward because there are no interconnections between the output of a processing element and the input of a node in the same layer or in a preceding layer. Externally provided correct patterns are compared with the neural network's output during (supervised) training, and feedback is used to adjust the weights until the network has categorized all training patterns as correctly as possible (the error tolerance is set in advance).

Starting with the output layer, errors between network-generated actual output and the desired outputs are used to correct/adjust the weights for the connections between the neurons (see Figure 6.13). For any output neuron j, the error (delta) $= (Z_j - Y_j)\,(df/dx)$, where Z and Y are the desired and actual outputs, respectively. Using the sigmoid function, $f = [1 + \exp(-x)]^{-1}$, where x is proportional to the sum of the weighted inputs to the neuron, is an effective way to compute the output of a neuron in practice. With this function, the derivative of the sigmoid function $df/dx = f(1 - f)$ and of the error is a simple function of the desired and actual outputs. The factor $f(1 - f)$ is the logistic function, which serves to keep the error correction well bounded. The weight of each input to the j^{th} neuron is then changed in proportion to this calculated error. A more complicated expression can be derived to work backward in a similar way from the output neurons through the hidden layers to calculate the corrections to the associated weights of the inner neurons. This complicated method is an iterative approach to solving a nonlinear optimization problem that is very similar in meaning to the one characterizing multiple linear regression.

In backpropagation, the learning algorithm includes the following procedures:

1. Initialize weights with random values and set other parameters.
2. Read in the input vector and the desired output.
3. Compute the actual output via the calculations, working forward through the layers.
4. Compute the error.
5. Change the weights by working backward from the output layer through the hidden layers.

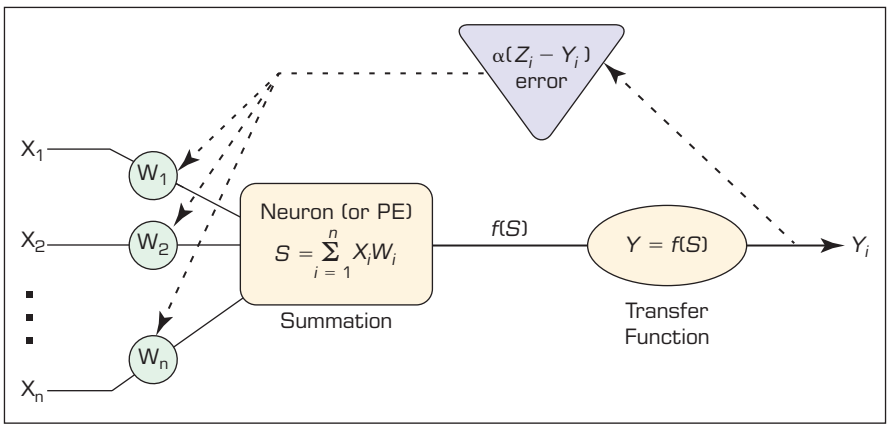


FIGURE 6.13 **Backpropagation of Error for a Single Neuron.**

This procedure is repeated for the entire set of input vectors until the desired output and the actual output agree within some predetermined tolerance. Given the calculation requirements for one iteration, training a large network can take a very long time; therefore, in one variation, a set of cases is run forward and an aggregated error is fed backward to speed the learning. Sometimes, depending on the initial random weights and network parameters, the network does not converge to a satisfactory performance level. When this is the case, new random weights must be generated, and the network parameters, or even its structure, may have to be modified before another attempt is made. Current research is aimed at developing algorithms and using parallel computers to improve this process. For example, genetic algorithms (GA) can be used to guide the selection of the network parameters to maximize the performance of the desired output. In fact, most commercial ANN software tools are now using GA to help users "optimize" the network parameters in a semiautomated manner.

A central concern in the training of any type of machine-learning model is **overfitting**. It happens when the trained model is highly fitted to the training data set but performs poorly with regard to external data sets. Overfitting causes serious issues with respect to the generalizability of the model. A large group of strategies known as *regularization* strategies is designed to prevent models from overfitting by making changes or defining constraints for the model parameters or the performance function.

In the classic ANN models of small size, a common regularization strategy to avoid overfitting is to assess the performance function for a separate validation data set as well as the training data set after each iteration. Whenever the performance stopped improving for the validation data, the training process would be stopped. Figure 6.14 shows a

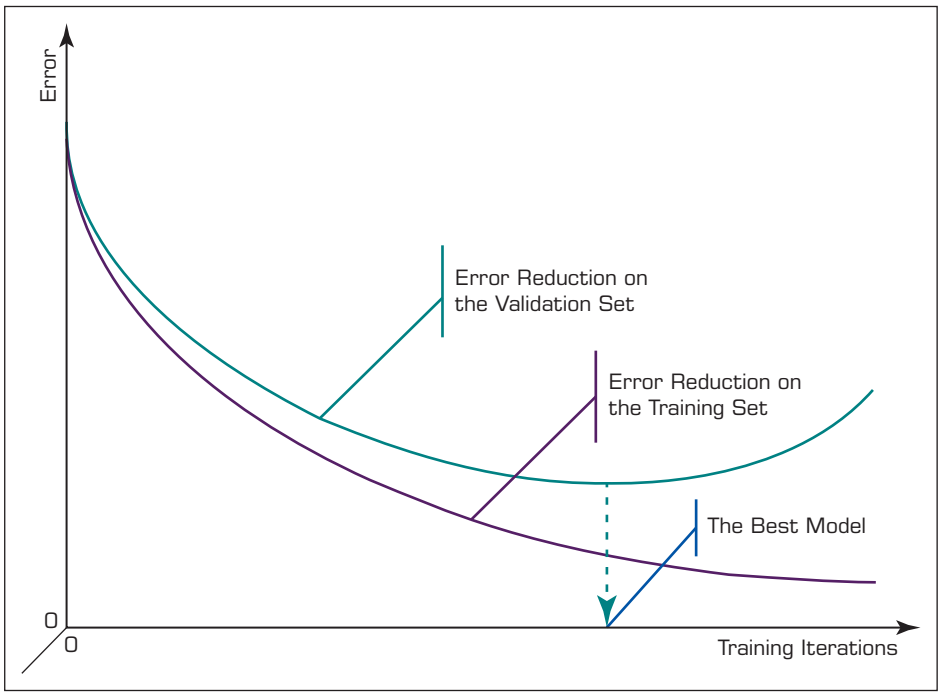


FIGURE 6.14 **Illustration of the Overfitting in ANN—Gradually Changing Error Rates in the Training and Validation Data Sets As the Number of Iterations Increases.**

typical graph of the error measure by the number of iterations of training. As shown, in the beginning, the error decreases in both training and validation data by running more and more iterations; but from a specific point (shown by the dashed line), the error starts increasing in the validation set while still decreasing in the training set. It means that beyond that number of iterations, the model becomes overfitted to the data set with which it is trained and cannot necessarily perform well when it is fed with some external data. That point actually represents the recommended number of iterations for training a given neural network.

Technology Insight 6.2 discusses some of the popular neural network software and offers some Web links to more comprehensive ANN-related software sites.

TECHNOLOGY INSIGHT 6.2 ANN Software

Many tools are available for developing neural networks (see this book's Web site and the resource lists at PC AI, **pcai.com**). Some of these tools function like software shells. They provide a set of standard architectures, learning algorithms, and parameters, along with the ability to manipulate the data. Some development tools can support several network paradigms and learning algorithms.

Neural network implementations are also available in most of the comprehensive predictive analytics and data mining tools, such as the SAS Enterprise Miner, IBM SPSS Modeler (formerly Clementine), and Statistica Data Miner. Weka, RapidMiner, Orange, and KNIME are open-source free data mining software tools that include neural network capabilities. These free tools can be downloaded from their respective Web sites; simple Internet searches on the names of these tools should lead you to the download pages. Also, most of the commercial software tools are available for download and use for evaluation purposes (usually they are limited on time of availability and/or functionality).

Many specialized neural network tools make the building and deployment of a neural network model an easier undertaking in practice. Any listing of such tools would be incomplete. Online resources such as Wikipedia (en.wikipedia.org/wiki/Artificial_neural_network), Google's or Yahoo!'s software directory, and the vendor listings on **pcai.com** are good places to locate the latest information on neural network software vendors. Some of the vendors that have been around for a while and have reported industrial applications of their neural network software include California Scientific (BrainMaker), NeuralWare, NeuroDimension Inc., Ward Systems Group (Neuroshell), and Megaputer. Again, the list can never be complete.

Some ANN development tools are spreadsheet add-ins. Most can read spreadsheet, database, and text files. Some are freeware or shareware. Some ANN systems have been developed in Java to run directly on the Web and are accessible through a Web browser interface. Other ANN products are designed to interface with expert systems as hybrid development products.

Developers may instead prefer to use more general programming languages, such as C, C#, C++, Java, and so on, readily available R and Python libraries, or spreadsheets to program the model, perform the calculations, and deploy the results. A common practice in this area is to use a library of ANN routines. Many ANN software providers and open-source platforms provide such programmable libraries. For example, hav.Software (**hav.com**) provides a library of C++ classes for implementing stand-alone or embedded feedforward, simple recurrent, and random-order recurrent neural networks. Computational software such as MATLAB also includes neural network–specific libraries.

SECTION 6.4 REVIEW QUESTIONS

1. List the nine steps in conducting a neural network project.
2. What are some of the design parameters for developing a neural network?
3. Draw and briefly explain the three-step process of learning in ANN.
4. How does backpropagation learning work?
5. What is overfitting in ANN learning? How does it happen, and how can it be mitigated?
6. Describe the different types of neural network software available today.

6.5 ILLUMINATING THE BLACK BOX OF ANN

Neural networks have been used as an effective tool for solving highly complex real-world problems in a wide range of application areas. Even though ANN have been proven to be superior predictors and/or cluster identifiers in many problem scenarios (compared to their traditional counterparts), in some applications, there exists an additional need to know "how the model does what it does." ANN are typically known as black boxes, capable of solving complex problems but lacking the explanation of their capabilities. This lack of transparency situation is commonly referred to as the "black-box" syndrome.

It is important to be able to explain a model's "inner being"; such an explanation offers assurance that the network has been properly trained and will behave as desired once deployed in a business analytics environment. Such a need to "look under the hood" might be attributable to a relatively small training set (as a result of the high cost of data acquisition) or a very high liability in case of a system error. One example of such an application is the deployment of airbags in vehicles. Here, both the cost of data acquisition (crashing vehicles) and the liability concerns (danger to human lives) are rather significant. Another representative example for the importance of explanation is loan-application processing. If an applicant is refused a loan, he or she has the right to know why. Having a prediction system that does a good job on differentiating good and bad applications may not be sufficient if it does not also provide the justification of its predictions.

A variety of techniques have been proposed for analysis and evaluation of trained neural networks. These techniques provide a clear interpretation of how a neural network does what it does; that is, specifically how (and to what extent) the individual inputs factor into the generation of specific network output. Sensitivity analysis has been the front-runner of the techniques proposed for shedding light into the black-box characterization of trained neural networks.

Sensitivity analysis is a method for extracting the cause-and-effect relationships among the inputs and the outputs of a trained neural network model. In the process of performing sensitivity analysis, the trained neural network's learning capability is disabled so that the network weights are not affected. The basic procedure behind sensitivity analysis is that the inputs to the network are systematically perturbed within the allowable value ranges, and the corresponding change in the output is recorded for each and every input variable (Principe et al., 2000). Figure 6.15 shows a graphical illustration of this process. The first input is varied between its mean plus and minus of a user-defined number of standard deviations (or for categorical variables, all of its possible values are used) while all other input variables are fixed at their respective means (or modes). The network output is computed for a user-defined number of steps above and below the mean. This process is repeated for each input. As a result, a report is generated to summarize the variation of each output with respect to the variation in each input. The generated report often contains a column plot (along with

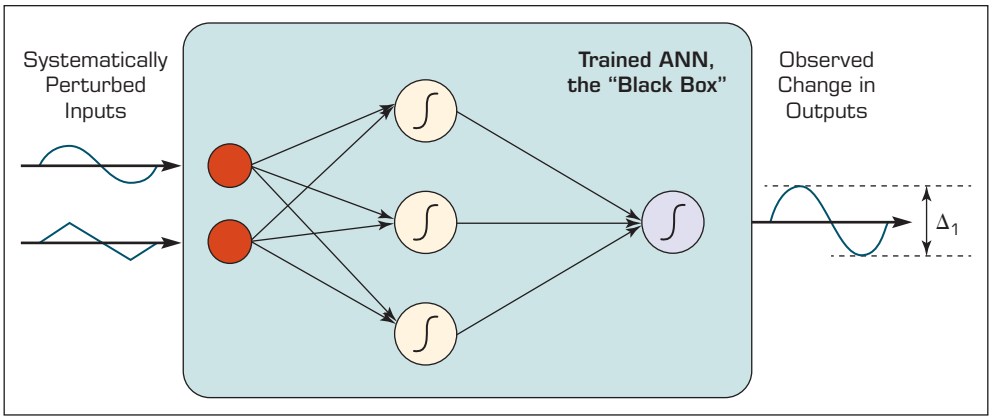


FIGURE 6.15 **A Figurative Illustration of Sensitivity Analysis on an ANN Model.**

numeric values presented on the *x*-axis), reporting the relative sensitivity values for each input variable. A representative example of sensitivity analysis on ANN models is provided in Application Case 6.4.

Application Case 6.4

Sensitivity Analysis Reveals Injury Severity Factors in Traffic Accidents

According to the National Highway Traffic Safety Administration (NHTSA), over 6 million traffic accidents claim more than 41,000 lives each year in the United States. Causes of accidents and related injury severity are of special interest to traffic safety researchers. Such research is aimed at reducing not only the number of accidents but also the severity of injury. One way to accomplish the latter is to identify the most profound factors that affect injury severity. Understanding the circumstances under which drivers and passengers are more likely to be severely injured (or killed) in a vehicle accident can help improve the overall driving safety situation. Factors that potentially elevate the risk of injury severity of vehicle occupants in the event of an accident include demographic and/or behavioral characteristics of the person (e.g., age, gender, seatbelt usage, use of drugs or alcohol while driving), environmental factors, and/or roadway conditions at the time of the accident (e.g., surface conditions, weather or light conditions, direction of the impact, vehicle orientation in the crash, occurrence of a rollover), as well as technical characteristics of the vehicle itself (e.g., age, body type).

In an exploratory data mining study, Delen et al. (2006) used a large sample of data—30,358 police-reported accident records obtained from the General Estimates System of NHTSA—to identify which factors become increasingly more important in escalating the probability of injury severity during a traffic crash. Accidents examined in this study included a geographically representative sample of multiple-vehicle collision accidents, single-vehicle fixed-object collisions, and single-vehicle noncollision (rollover) crashes.

Contrary to many of the previous studies conducted in this domain, which have primarily used regression-type generalized linear models where the functional relationships between injury severity and crash-related factors are assumed to be linear (which is an oversimplification of the reality in most real-world situations), Delen and his colleagues (2006) decided to go in a different direction. Because ANN are known to be superior in capturing highly nonlinear complex relationships between the predictor variables (crash factors) and the target variable (severity level of the injuries), they decided to use a series of ANN models to estimate the significance of the crash factors on the level of injury severity sustained by the driver.

(*Continued*)

Application Case 6.4 (Continued)

From a methodological standpoint, Delen et al. (2006) followed a two-step process. In the first step, they developed a series of prediction models (one for each injury severity level) to capture the in-depth relationships between the crash-related factors and a specific level of injury severity. In the second step, they conducted sensitivity analysis on the trained neural network models to identify the prioritized importance of crash-related factors as they relate to different injury severity levels. In the formulation of the study, the five-class prediction problem was decomposed into a number of binary classification models to obtain the granularity of information needed to identify the "true" cause-and-effect relationships between the crash-related factors and different levels of injury severity. As shown in Figure 6.16, eight different neural network models have been developed and used in the sensitivity analysis to identify the key determinants of increased injury severity levels.

The results revealed considerable differences among the models built for different injury severity levels. This implies that the most influential factors in prediction models highly depend on the level of injury severity. For example, the study revealed that the variable seatbelt use was the most important determinant for predicting higher levels of injury severity (such as incapacitating injury or fatality), but it was one of the least significant predictors for lower levels of injury severity (such as non-incapacitating injury and minor injury). Another interesting finding involved gender: The driver's gender was among the significant predictors for lower levels of injury severity, but it was not among the significant factors for higher levels of injury severity, indicating that more serious injuries do not depend on the driver being a male or a female. Another interesting and somewhat intuitive finding of the study indicated that age becomes an increasingly more significant factor as the level of injury severity increases, implying that older people are more likely to incur severe injuries (and fatalities) in serious vehicle crashes than younger people.

Questions for Case 6.4

1. How does sensitivity analysis shed light on the black box (i.e., neural networks)?
2. Why would someone choose to use a black-box tool such as neural networks over theoretically sound, mostly transparent statistical tools like logistic regression?
3. In this case, how did neural networks and sensitivity analysis help identify injury-severity factors in traffic accidents?

Sources: Delen, D., R. Sharda, & M. Bessonov. (2006). "Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks." *Accident Analysis and Prevention, 38*(3), pp. 434–444; Delen, D., L. Tomak, K. Topuz, & E. Eryarsoy (2017). "Investigating Injury Severity Risk Factors in Automobile Crashes with Predictive Analytics and Sensitivity Analysis Methods." *Journal of Transport & Health, 4*, pp. 118–131.

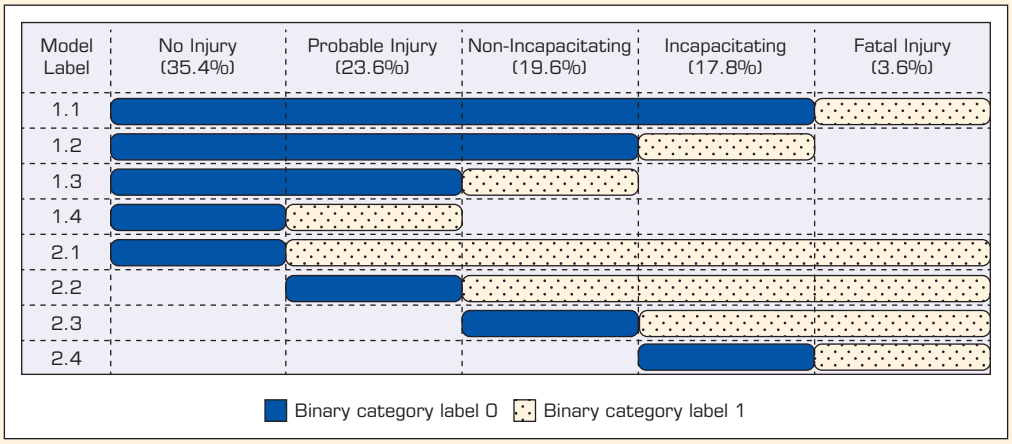


FIGURE 6.16 Graphical Representation of the Eight Binary ANN Model Configurations.

SECTION 6.5 REVIEW QUESTIONS

1. What is the so-called black-box syndrome?
2. Why is it important to be able to explain an ANN's model structure?
3. How does sensitivity analysis work in ANN?
4. Search the Internet to find other methods to explain ANN methods. Report the results.

6.6 DEEP NEURAL NETWORKS

Until recently (before the advent of deep learning phenomenon), most neural network applications involved network architectures with only a few hidden layers and a limited number of neurons in each layer. Even in relatively complex business applications of neural networks, the number of neurons in networks hardly exceeded a few thousands. In fact, the processing capability of computers at the time was such a limiting factor that central processing units (CPU) were hardly able to run networks involving more than a couple of layers in a reasonable time. In recent years, development of graphics processing units (GPUs) along with the associated programming languages (e.g., CUDA by NVIDIA) that enable people to use them for data analysis purposes has led to more advanced applications of neural networks. GPU technology has enabled us to successfully run neural networks with over a million neurons. These larger networks are able to go deeper into the data features and extract more sophisticated patterns that could not be detected otherwise.

While deep networks can handle a considerably larger number of input variables, they also need relatively larger data sets to be trained satisfactorily; using small data sets for training deep networks typically leads to overfitting of the model to the training data and poor and unreliable results in case of applying to external data. Thanks to the Internet- and Internet of Things (IoT)-based data-capturing tools and technologies, larger data sets are now available in many application domains for deeper neural network training.

The input to a regular ANN model is typically an array of size R × 1, where R is the number of input variables. In the deep networks, however, we are able to use *tensors* (i.e., N-dimensional arrays) as input. For example, in image recognition networks, each input (i.e., image) can be represented by a matrix indicating the color codes used in the image pixels; or for video processing purposes, each video can be represented by several matrices (i.e., a 3D tensor), each representing an image involved in the video. In other words, tensors provide us with the ability to include additional dimensions (e.g., time, location) in analyzing the data sets.

Except for these general differences, the different types of deep networks involve various modifications to the architecture of standard neural networks that equip them with distinct capabilities of dealing with particular data types for advanced purposes. In the following section, we discuss some of these special network types and their characteristics.

Feedforward Multilayer Perceptron (MLP)-Type Deep Networks

MLP deep networks, also known as *deep feedforward networks*, are the most general type of deep networks. These networks are simply large-scale neural networks that can contain many layers of neurons and handle tensors as their input. The types and characteristics of the network elements (i.e., weight functions, transfer functions) are pretty much the same as in the standard ANN models. These models are called *feedforward* because the flow of information that goes through them is always forwarding and no feedback connections (i.e., connections in which outputs of a model are fed back to itself) are allowed. The neural networks in which feedback connections are allowed are called *recurrent neural networks (RNN)*. General RNN architectures, as well as a specific variation of RNNs called *long short-term memory networks*, are discussed in later sections of this chapter.

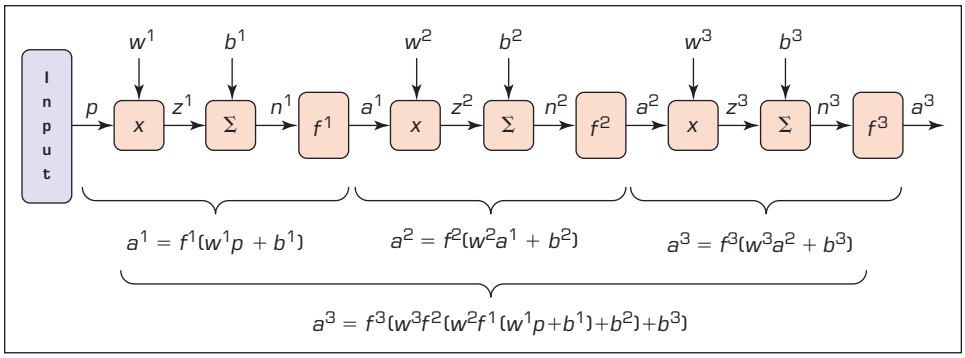


FIGURE 6.17 **Vector Representation of the First Three Layers in a Typical MLP Network.**

Generally, a sequential order of layers has to be held between the input and the output layers in the MLP-type network architecture. This means that the input vector has to pass through all layers sequentially and cannot skip any of them; moreover, it cannot be directly connected to any layer except for the very first one; the output of each layer is the input to the subsequent layer. Figure 6.17 demonstrates a vector representation of the first three layers of a typical MLP network. As shown, there is only one vector going into each layer, which is either the original input vector (p for the first layer) or the output vector from the previous hidden layer in the network architecture (a^{i-1} for the i^{th} layer). There are, however, some special variations of MLP network architectures designed for specialized purposes in which these principles can be violated.

Impact of Random Weights in Deep MLP

Optimization of the performance (loss) function in many real applications of deep MLPs is a challenging issue. The problem is that applying the common gradient-based training algorithms with random initialization of weights and biases that is very efficient for finding the optimal set of parameters in shallow neural networks most of the time could lead to getting stuck in the locally optimal solutions rather than catching the global optimum values for the parameters. As the depth of network increases, chances of reaching a global optimum using random initializations with the gradient-based algorithms decrease. In such cases, usually pretraining the network parameters using some *unsupervised* deep learning methods such as **deep belief networks (DBNs)** can be helpful (Hinton, Osindero, and Teh, 2006). DBNs are a type of a large class of deep neural networks called *generative models*. Introduction of DBNs in 2006 is considered as the beginning of the current deep learning renaissance (Goodfellow et al., 2016), since prior to that, deep models were considered too difficult to optimize. In fact, the primary application of DBNs today is to improve classification models by pretraining of their parameters.

Using these unsupervised learning methods, we can train the MLP layers, one at a time, starting from the first layer, and use the output of each layer as the input to the subsequent layer and initialize that layer with an unsupervised learning algorithm. At the end, we will have a set of initialized values for the parameters across the whole network. Those pretrained parameters, instead of random initialized parameters, then can be used as the initial values in the supervised learning of the MLP. This pretraining procedure has been shown to cause significant improvements to the deep classification applications. Figure 6.18 illustrates the classification errors that resulted from training a deep MLP network with (blue circles) and without (black triangles) pretraining of parameters (Bengio, 2009). In this example, the blue line represents the observed error rates of testing a classification model (on 1,000 heldout examples) trained using a purely supervised approach with 10 million examples,

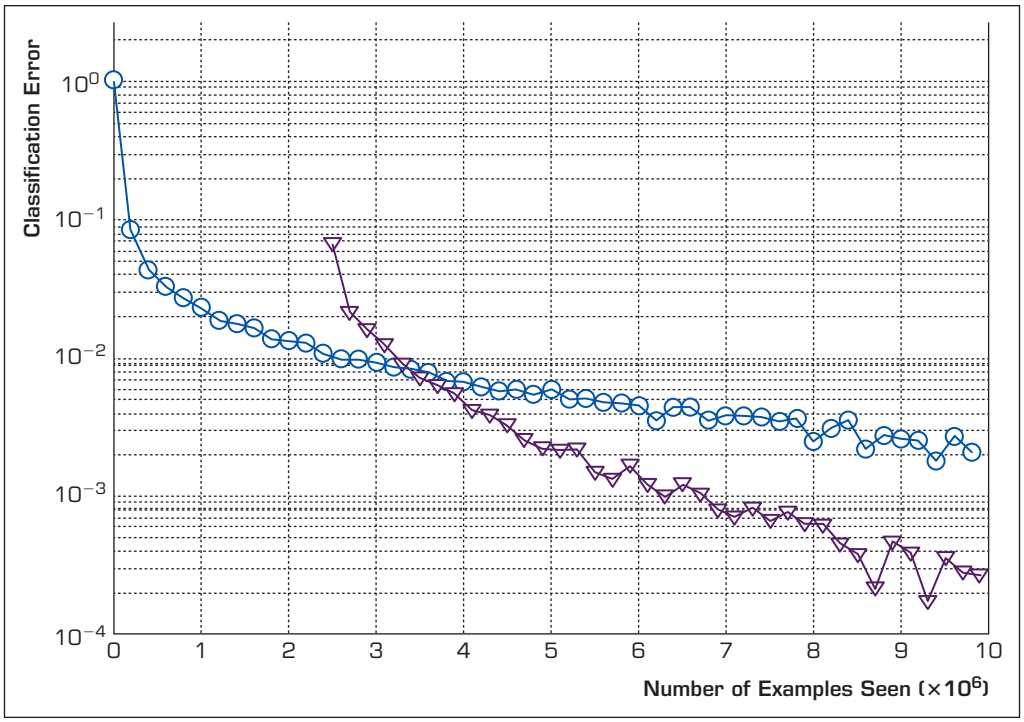


FIGURE 6.18 **The Effect of Pretraining Network Parameters on Improving Results of a Classification-Type Deep Neural Network.**

whereas the black line indicates the error rates on the same testing data set when 2.5 million examples were initially used for unsupervised training of network parameters (using DBN) and then the other 7.5 million examples along with the initialized parameters were used to train a supervised classification model. The diagrams clearly show a significant improvement in terms of the classification error rate in the model pretrained by a deep belief network.

More Hidden Layers versus More Neurons?

An important question regarding the deep MLP models is "Would it make sense (and produce better results) to restructure such networks with only a few layers, but many neurons in each?" In other words, the question is why do we need deep MLP networks with many layers when we can include the same number of neurons in just a few layers (i.e., wide networks instead of deep networks). According to the universal approximation theorem (Cybenko, 1989; Hornik, 1991), a sufficiently large single-layer MLP network will be able to approximate any function. Although theoretically founded, such a layer with many neurons may be prohibitively large and hence may fail to learn the underlying patterns correctly. A deeper network can reduce the number of neurons required at each layer and hence decrease the generalization error. Whereas theoretically it is still an open research question, practically using more layers in a network seems to be more effective and computationally more efficient than using many neurons in a few layers.

Like typical artificial neural networks, multilayer perceptron networks can also be used for various prediction, classification, and clustering purposes. Especially when a large number of input variables are involved or in cases that the nature of input has to be an *N*-dimensional array, a deep multilayer network design needs to be employed.

Application Case 6.5 provides an excellent case for the use of advanced analytics to better manage traffic flows in crowded cities.

Application Case 6.5

Georgia DOT Variable Speed Limit Analytics Help Solve Traffic Congestions

The Background

When the Georgia Department of Transportation (GDOT) wanted to optimize the use of Big Data and advanced analytics to gain insight into transportation, it worked with Teradata to develop a proof of concept evaluation of GDOT's variable speed limit (VSL) pilot project.

The VSL concept has been adopted in many parts of the world, but it is still relatively new in the United States. As GDOT explains,

> VSL are speed limits that change based on road, traffic, and weather conditions. Electronic signs slow down traffic ahead of congestion or bad weather to smooth out flow, diminish stop-and-go conditions, and reduce crashes. This low-cost, cutting edge technology alerts drivers in real time to speed changes due to conditions down the road. More consistent speeds improve safety by helping to prevent rear-end and lane changing collisions due to sudden stops.

Quantifying the customer service, safety, and efficiency benefits of VSL is extremely important to GDOT. This fits within a wider need to understand the effects of investments in intelligent transportation systems as well as other transportation systems and infrastructures.

VSL Pilot Project on I-285 in Atlanta

GDOT conducted a VSL pilot project on the northern half, or "top end," of I-285 that encircles Atlanta. This 36-mile stretch of highway was equipped with 88 electronic speed limit signs that adjusted speed limits in 10 mph increments from 65 miles per hour (mph) to the minimum of 35 mph. The objectives were twofold:

1. Analyze speeds on the highway before versus after implementation of VSL.
2. Measure the impact of VSL on driving conditions.

To obtain an initial view of the traffic, the Teradata data science solution identified the locations and durations of "persistent slowdowns." If highway speeds are above "reference speed," then traffic is considered freely flowing. Falling below the reference speed at any point on the highway is considered a slowdown. When slowdowns persist across multiple consecutive minutes, a persistent slowdown can be defined.

By creating an analytic definition of slowdowns, it is possible to convert voluminous and highly variable speed data into patterns to support closer investigation. The early analyses of the data revealed that the clockwise and counterclockwise directions of the same highway may show significantly different frequency and duration of slowdowns. To better understand how slowdowns affect highway traffic, it is useful to take our new definition and zoom in on a specific situation. Figure 6.19 shows a specific but typical Atlanta afternoon on I-285, at a section of highway where traffic is moving clockwise, from west to east, between mile markers MM10 in the west to the east end at MM46.

The first significant slowdown occurred at 3:00 P.M. near MM32. The size of the circles represents duration (measured in minutes). The slowdown at MM32 was nearly four hours long. As the slowdown "persisted," traffic speed diminished behind it. The slowdown formed on MM32 became a bottleneck that caused traffic behind it to slow down as well. The "comet trail" of backed-up traffic at the top left of Figure 6.20 illustrates the sequential formation of slowdowns at MM32 and then farther west, each starting later in the afternoon and not lasting as long.

Measuring Highway Speed Variability

The patterns of slowdowns on the highway as well as their different timings and locations led us to question their impact on drivers. If VSL could help drivers better anticipate the stop-and-go nature of the slowdowns, then being able to quantify the impact would be of interest to GDOT. GDOT was particularly concerned about what happens when a driver first encounters a slowdown. "While we do not know what causes the slowdown, we do know that drivers have made speed adjustments. If the slowdown was caused by an accident, then the speed reduction could be quite sudden; alternatively, if the slowdown was just caused by growing volumes of traffic, then the speed reduction might be much more gradual."

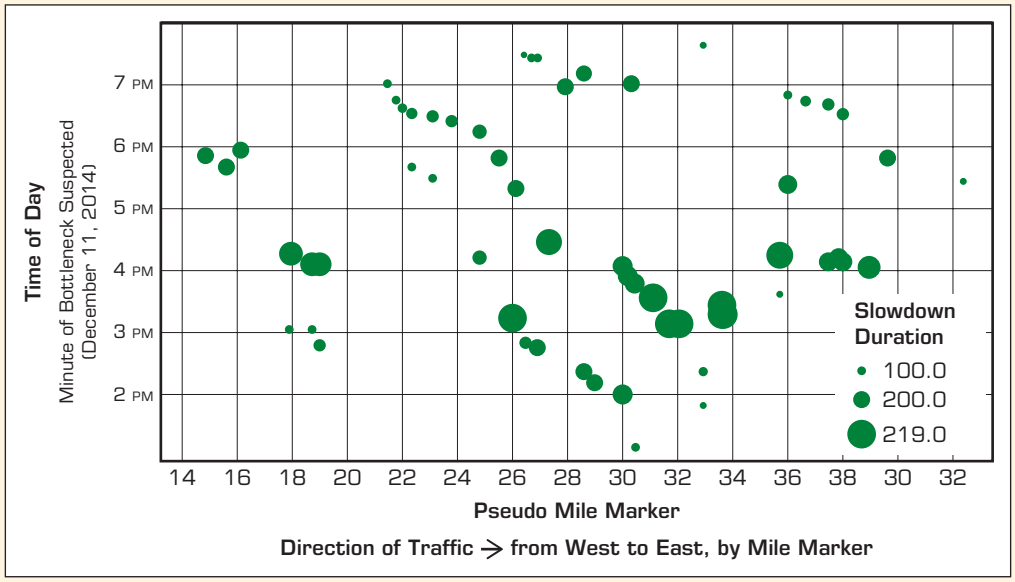


FIGURE 6.19 **Traffic Moving Clockwise during the Afternoon.**

Identifying Bottlenecks and Traffic Turbulence

A bottleneck starts as a slowdown at a particular location. Something like a "pinch point" occurs on the highway. Then, over a period of time, traffic slows down behind the original pinch point. A bottleneck is a length of highway where traffic falls below 60 percent of reverence speed and can stay at that level for miles. Figure 6.20 shows a conceptual representation of a bottleneck.

While bottlenecks are initiated by a pitch point, or slowdown, that forms the head of the queue, it is the end of the queue that is the most interesting. The area at the back of a queue is where traffic encounters a transition from free flow to slowly

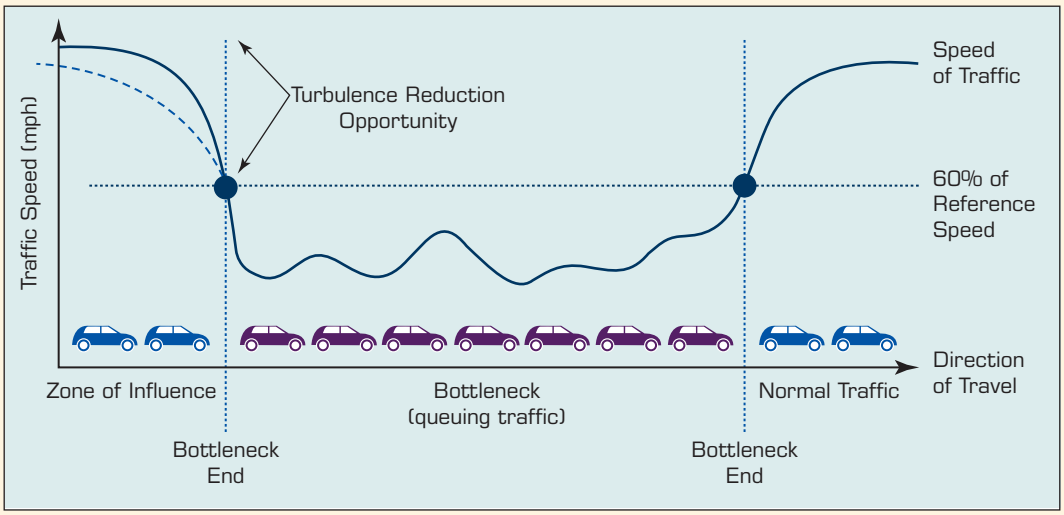


FIGURE 6.20 **Graphical Depiction of a Bottleneck on a Highway.**

(Continued)

Application Case 6.5 (Continued)

moving congested conditions. In the worst conditions, the end of the queue can experience a rapid transition. Drivers moving at highway speed may unexpectedly encounter slower traffic. This condition is ripe for accidents and is the place where VSL can deliver real value.

Powerful New Insight on Highway Congestion

The availability of new Big Data sources that describe the "ground truth" of traffic conditions on highways provides rich new opportunities for developing and analyzing highway performance metrics. Using just a single data source on detailed highway speeds, we produced two new and distinctive metrics using Teradata advanced data science capabilities.

First, by defining and measuring persistent slowdowns, we helped traffic engineers understand the frequency and duration of slow speed locations on a highway. The distinction of measuring a persistent slowdown versus a fleeting one is uniquely challenging and requires data science. It provides the ability to compare the number, duration, and location of slowdowns in a way that is more informative and compelling than simple averages, variances, and outliers in highway speeds.

The second metric was the ability to measure turbulence caused by bottlenecks. By identifying where bottlenecks occur and then narrowing in on their very critical zones of influence, we can make measurements of speeds and traffic deceleration turbulence within those zones. Data science and analytics capabilities demonstrated reduced turbulence when VSL is active in the critical zone of a bottleneck.

There is much more that could be explored within this context. For example, it is natural to assume that because most traffic is on the road during rush hours, VSL provides the most benefits during these high-traffic periods. However, the opposite may be true, which could provide a very important benefit of the VSL program.

Although this project was small in size and was just a proof of concept, a combination of similar projects beyond just transportation under the name of "smart cities" is underway around the United States and abroad. The goal is to use a variety of data from sensors to multimedia, rare event reports to satellite images along with advanced analytics that include deep learning and cognitive computing to transform the dynamic nature of cities toward better to best for all stakeholders.

Questions for Case 6.5

1. What was the nature of the problems that GDOT was trying to solve with data science?
2. What type of data do you think was used for the analytics?
3. What were the data science metrics developed in this pilot project? Can you think of other metrics that can be used in this context?

Source: Teradata Case Study. "Georgia DOT Variable Speed Limit Analytics Help Solve Traffic Congestion." **https:// www.teradata.com/Resources/Case-Studies/Georgia-DOT-Variable-Speed-Limit-Analytics** (accessed July 2018); "Georgia DOT Variable Speed Limits." **www.dot.ga.gov/ DriveSmart/SafetyOperation/Pages/VSL.aspx** (accessed August 2018).Used with permission from Teradata.

In the next section, we discuss a very popular variation of deep MLP architecture called **convolutional neural network (CNN)** specifically designed for computer vision applications (e.g., image recognition, handwritten text processing).

SECTION 6.6 REVIEW QUESTIONS

1. What is meant by "deep" in deep neural networks? Compare deep neural networks to shallow neural networks.
2. What is GPU? How does it relate to deep neural networks?
3. How does a feedforward multilayer perceptron-type deep network work?

4. Comment on the impact of random weights in developing deep MLP.
5. Which strategy is better: more hidden layers versus more neurons?

6.7 CONVOLUTIONAL NEURAL NETWORKS

CNNs (LeCun et al., 1989) are among the most popular types of deep learning methods. CNNs are in essence variations of the deep MLP architecture, initially designed for computer vision applications (e.g., image processing, video processing, text recognition) but are also applicable to nonimage data sets.

The main characteristic of the convolutional networks is having at least one layer involving a *convolution weight function* instead of general matrix multiplication. Figure 6.21 illustrates a typical convolutional unit.

Convolution, typically shown by the $\circledast$ symbol, is a linear operation that essentially aims at extracting simple patterns from sophisticated data patterns. For instance, in processing an image containing several objects and colors, convolution functions can extract simple patterns like the existence of horizontal or vertical lines or edges in different parts of the picture. We discuss convolution functions in more detail in the next section.

A layer containing a convolution function in a CNN is called a *convolution layer*. This layer is often followed by a **pooling** (a.k.a. *subsampling*) layer. Pooling layers are in charge of consolidating the large tensors to one with a smaller size and reducing the number of model parameters while keeping their important features. Different types of pooling layers are also discussed in the following sections.

Convolution Function

In the description of MLP networks, it was said that the weight function is generally a matrix manipulation function that multiplies the weight vector into the input vector to produce the output vector in each layer. Having a very large input vector/tensor, which is the case in most deep learning applications, we need a large number of weight parameters so that each single input to each neuron could be assigned a single weight parameter. For instance, in an image-processing task using a neural network for images of size 150 × 150 pixels, each input matrix will contain 22,500 (i.e., 150 times 150) integers, each of which should be assigned its own weight parameter per each neuron it goes into throughout the network. Therefore, having even only a single layer requires thousands of weight parameters to be defined and trained. As one might guess, this fact would dramatically increase the required time and processing power to train a network, since in each training iteration, all of those weight parameters have to be updated by the SGD algorithm. The solution to this problem is the convolution function.

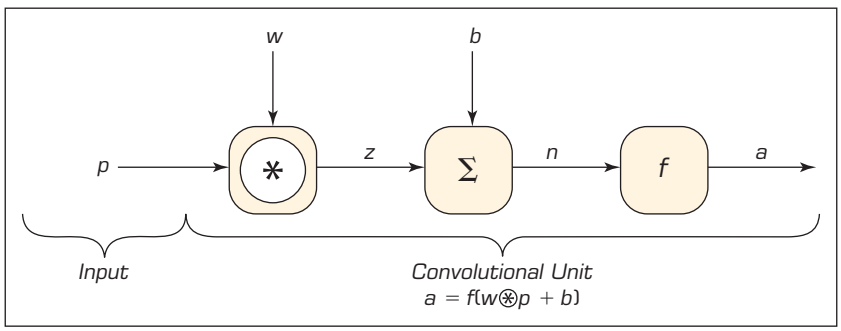


FIGURE 6.21 Typical Convolutional Network Unit.

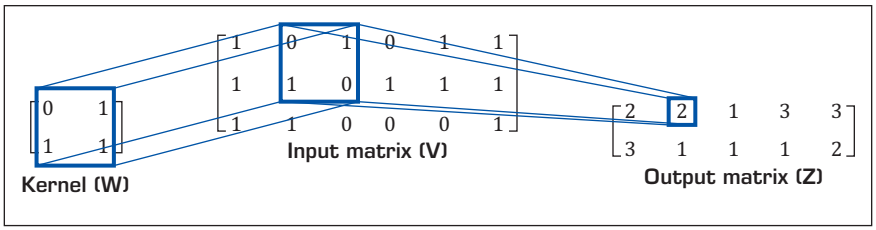


FIGURE 6.22 **Convolution of a 2 × 2 Kernel by a 3 × 6 Input Matrix.**

The convolution function can be thought of as a trick to address the issue defined in the previous paragraph. The trick is called *parameter sharing*, which in addition to computational efficiency provides additional benefits. Specifically, in a convolution layer, instead of having a weight for each input, there is a set of weights referred to as the *convolution kernel* or *filter*, which is shared between inputs and moves around the input matrix to produce the outputs. The kernel is typically represented as a small matrix of size $W_{r\times c}$; for a given input matrix V, then, the convolution function can be stated as:

$$z_{i,j} = \sum_{k=1}^{r}\sum_{l=1}^{c} w_{k,l} v_{i+k-1,\, j+l-1}$$

For example, assume that the input matrix to a layer and the convolution kernel is

$$V = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad W = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

Figure 6.22 illustrates how the convolution output can be computed. As shown, each element of the output matrix results from summing up the one-by-one point multiplications of the kernel elements into a corresponding r × c (in this example, 2 × 2 because the kernel is 2 × 2) subset of the input matrix elements. So, in the example shown, the element at the second column of the first row of the output matrix is in fact 0(0) + 1(1) + 1(1) + 1(0) = 2.

It can be seen that the magnitude of each element in the output matrix directly depends on how the matched kernel (with the 2 × 2 matrix) and the input matrix are involved in calculation of that element. For example, the element at the fourth column of the first row of the output matrix is the result of convoluting the kernel by a part of the input matrix, which is exactly the same as the kernel (shown in Figure 6.23). This suggests that by applying the convolution operation, we actually are converting the input matrix into an output in which the parts that have a particular feature (reflected by the kernel) are placed in the square box.

This characteristic of convolution functions is especially useful in practical image-processing applications. For instance, if the input matrix represents the pixels of an image,

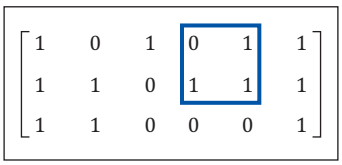


FIGURE 6.23 **The Output of Convolution Operation Is Maximized When the Kernel Exactly Matches the Part of Input Matrix That Is Being Convoluted by.**

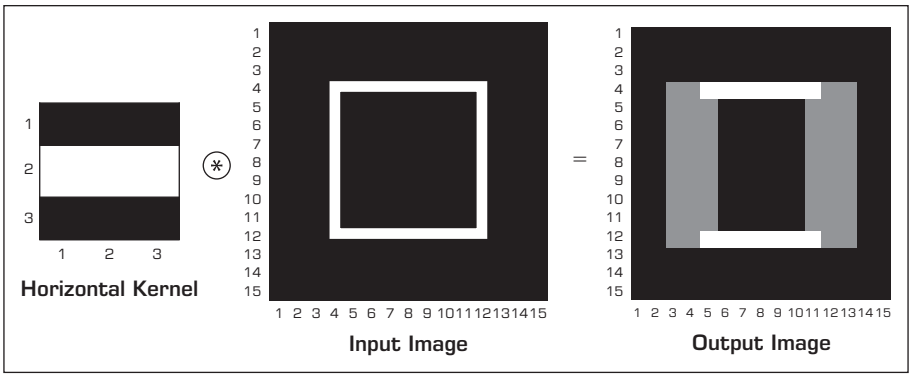


FIGURE 6.24 **Example of Using Convolution for Extracting Features (Horizontal Lines in This Example) from Images.**

a particular kernel representing a specific shape (e.g., a diagonal line) may be convoluted into that image to extract parts of the image involving that specific shape. Figure 6.24, for example, shows the result of applying a 3×3 horizontal line kernel to a 15×15 image of a square.

Clearly, the horizontal kernel produces an output in which the location of horizontal lines (as a feature) in the original input image is identified.

Convolution using a kernel of size $r \times c$ will reduce the number of rows and columns in the output by $r - 1$ and $c - 1$, respectively. In the recent case, for example, using a 2×2 kernel for convolution, the output matrix has 1 row and 1 column less than the input matrix. To prevent this change of size, we can *pad* the outside of the input matrix with zeros before convolving, that is, to add $r - 1$ rows and $c - 1$ columns of zeros to the input matrix. On the other hand, if we want the output matrix to be even smaller, we can have the kernel to take larger *strides*, or kernel movements. Normally, the kernel is moved one step at a time (i.e., stride $= 1$) when performing the convolution. By increasing this stride to 2, the size of the output matrix is reduced by a factor of 2.

Although the main benefit of employing convolution in the deep networks is parameter sharing, which effectively reduces the required time and processing power to train the network by reducing the number of weight parameters, it involves some other benefits as well. A convolution layer in a network will have a property called *equivariance* for translation purposes (Goodfellow et al., 2016). It simply means that any changes in the input will lead to a change in the output in the same way. For instance, moving an object in the input image by 10 pixels in a particular direction will lead to moving its representation in the output image by 10 pixels in the same direction. Apart from image-processing applications, this feature is especially useful for analyzing time-series data using convolutional networks where convolution can produce a kind of timeline that shows when each feature appears in the input.

It should be noted that in almost all of the practical applications of convolutional networks, many convolution operations are used in parallel to extract various kinds of features from the data, because a single feature is hardly enough to fully describe the inputs for the classification or recognition purposes. Also, as noted before, in most real-world applications, we have to represent the inputs as multi-dimensional tensors. For instance, in the processing of color images as opposed to gray scale pictures, instead of having 2D tensors (i.e., matrices) that represent the color of pixels (i.e., black or white), one will have to use 3D tensors because each pixel should be defined using the intensity of red, blue, and green colors.

Pooling

Most of the times, a convolution layer is followed by another layer known as the *pooling* (a.k.a. *subsampling*) layer. The purpose of a pooling layer is to consolidate elements in the input matrix to produce a smaller output matrix while maintaining the important features. Normally, a pooling function involves an r × c consolidation window (similar to a kernel in the convolution function) that moves around the input matrix and in each move calculates some summary statistics of the elements involved in the consolidation window so that it can be put in the output image. For example, a particular type of pooling function called *average pooling* takes the average of the input matrix elements involved in the consolidation window and puts that average value as an element of the output matrix in the corresponding location. Similarly, the *max pooling* function (Zhou et al.) takes the maximum of the values in the window as the output element. Unlike convolution, for the pooling function, given the size of the consolidation window (i.e., r and c), stride should be carefully selected so that there would be no overlaps in the consolidations. The pooling operation using an r × c consolidation window reduces the number of rows and columns of the input matrix by a factor of r and c, respectively. For example, using a 3 × 3 consolidation window, a 15 × 15 matrix will be consolidated to a 5 × 5 matrix.

Pooling, in addition to reducing the number of parameters, is especially useful in the image-processing applications of deep learning in which the critical task is to determine whether a feature (e.g., a particular animal) is present in an image while the exact spatial location of the same in the picture is not important. However, if the location of features is important in a particular context, applying a pooling function could potentially be misleading.

You can think of pooling as an operation that summarizes large inputs whose features are already extracted by the convolution layer and shows us just the important parts (i.e., features) in each small neighborhood in the input space. For instance, in the case of the image-processing example shown in Figure 6.24, if we place a max pooling layer after the convolution layer using a 3 × 3 consolidation window, the output will be like what is shown in Figure 6.25. As shown, the 15 × 15 already convoluted image is consolidated in a 5 × 5 image while the main features (i.e., horizontal lines) are maintained therein.

Sometimes pooling is used just to modify the size of matrices coming from the previous layer and convert them to a specified size required by the following layer in the network.

There are various types of pooling operations such as max pooling, average pooling, the L^2 norm of a rectangular neighborhood, and weighted average pooling. The

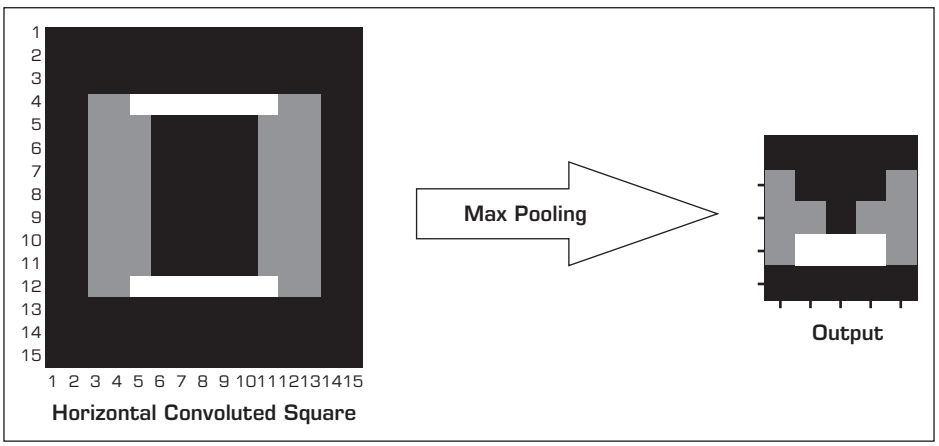


FIGURE 6.25 An Example of Applying Max Pooling on an Output Image to Reduce Its Size.

choice of proper pooling operation as well as the decision to include a pooling layer in the network at all depends highly on the context and properties of the problem that the network is solving. There are some guidelines in the literature to help the network designers in making such decisions (Boureau et al., 2011; Boureau, Ponce, and LeCun, 2010; Scherer, Müller, and Behnke, 2010).

Image Processing Using Convolutional Networks

Real applications of deep learning in general and CNNs in particular highly depend on the availability of large, annotated data sets. Theoretically, CNNs can be applied to many practical problems, and today there are many large and feature-rich databases for such applications available. Nevertheless, the biggest challenge is that in supervised learning applications, one needs an already annotated (i.e., labeled) data set to train the model before we can use it for prediction/identification of other unknown cases. Whereas extracting features of data sets using CNN layers is an unsupervised task, the extracted features will not be of much use without having labeled cases to develop a classification network in a supervised learning fashion. That is why image classification networks traditionally involve two pipelines: visual feature extraction and image classification.

ImageNet (**http://www.image-net.org**) is an ongoing research project that provides researchers with a large database of images, each linked to a set of synonym words (known as *synset*) from WordNet (a word hierarchy database). Each *synset* represents a particular concept in the WordNet. Currently, WordNet includes more than 100,000 synsets, each of which is supposed to be illustrated by an average of 1,000 images in the ImageNet. ImageNet is a huge database for developing image processing–type deep networks. It contains more than 15 million labeled images in 22,000 categories. Because of its sheer size and proper categorization, ImageNet is by far the most widely used benchmarking data set to assess the efficiency and accuracy of deep networks designed by deep learning researchers.

One of the first convolutional networks designed for image classification using the ImageNet data set was AlexNet (Krizhevsky, Sutskever, and Hinton, 2012). It was composed of five convolution layers followed by three fully connected (a.k.a. dense) layers (see Figure 6.26 for a schematic representation of AlexNet). One of the contributions of this relatively simple architecture that made its training remarkably faster and computationally efficient was the use of rectified linear unit (ReLu) transfer functions in the convolution layers instead of the traditional sigmoid functions. By doing so, the designers

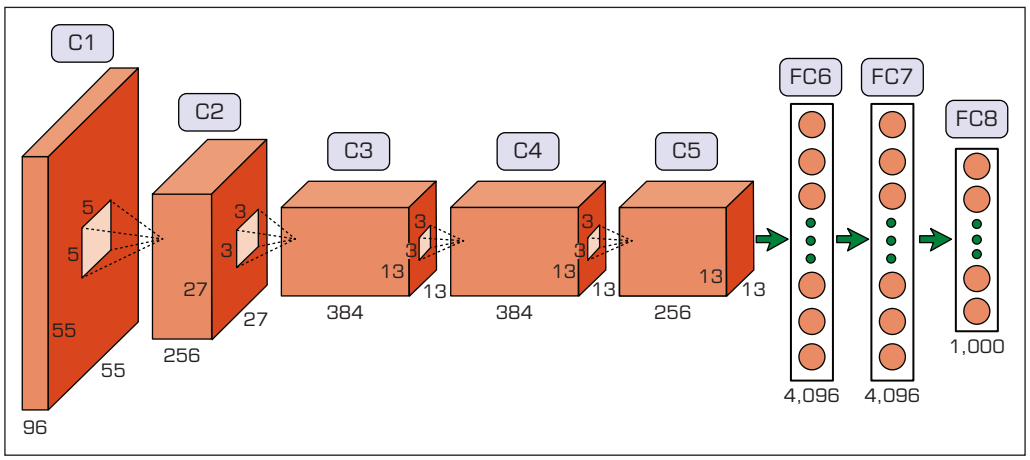


FIGURE 6.26 **Architecture of AlexNet, a Convolutional Network for Image Classification.**

addressed the issue called the *vanishing gradient problem* caused by very small derivatives of sigmoid functions in some regions of the images. The other important contribution of this network that has a dramatic role in improving the efficiency of deep networks was the introduction of the concept of dropout layers to the CNNs as a regularization technique to reduce overfitting. A dropout layer typically comes after the fully connected layers and applies a random probability to the neurons to switch off some of them and make the network sparser.

In the recent years, in addition to a large number of data scientists who showcase their deep learning capabilities, a number of well-known industry-leading companies such as Microsoft, Google, and Facebook have participated in the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC). The goal in the ILSVRC classification task is to design and train networks that are capable of classifying 1.2 million input images into one of the 1,000 image categories. For instance, GoogLeNet (a.k.a. Inception), a deep convolutional network architecture designed by Google researchers, was the winning architecture of ILSVRC 2014 with a 22-layer network and only a 6.66 percent classification error rate, only slightly (5.1%) worse than the human-level classification error (Russakovsky et al., 2015). The main contribution of the GoogLeNet architecture was to introduce a module called *Inception*. The idea of Inception is that because one would have no idea of the size of convolution kernel that would perform best on a particular data set, it is better to include multiple convolutions and let the network decide which one to use. Therefore, as shown in Figure 6.27, in each convolution layer, the data coming from the previous layer is passed through multiple types of convolution and the outputs are concatenated before going to the next layer. Such architecture allows the model to take into account both local features via smaller convolutions and high abstracted features via larger ones.

Google recently launched a new service, Google Lens, that uses deep learning artificial neural network algorithms (along with other AI techniques) to deliver information about the images captured by users from their nearby objects. This involves identifying the objects, products, plants, animals, and locations and providing information about them on the Internet. Some other features of this service are the capability of saving

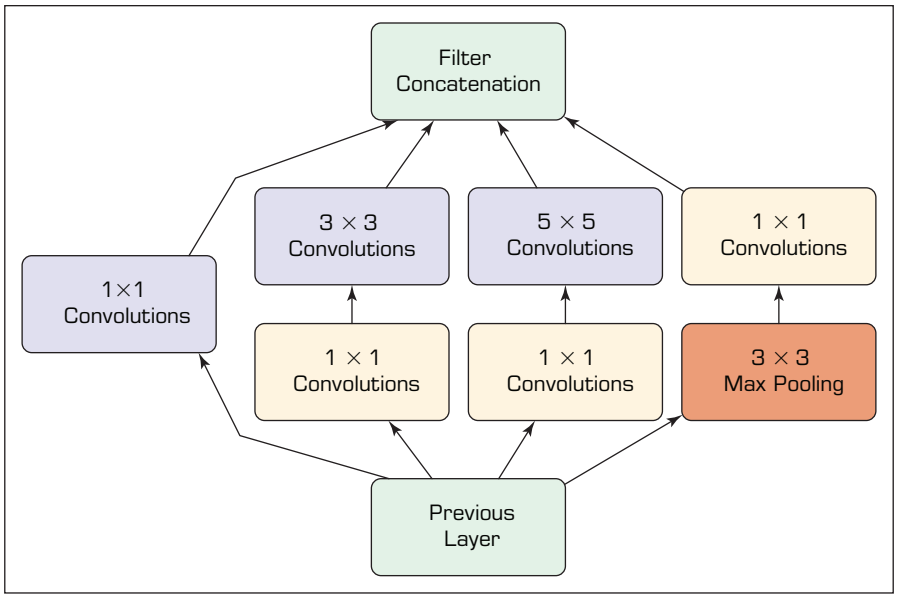


FIGURE 6.27 **Conceptual Representation of the Inception Feature in GoogLeNet.**

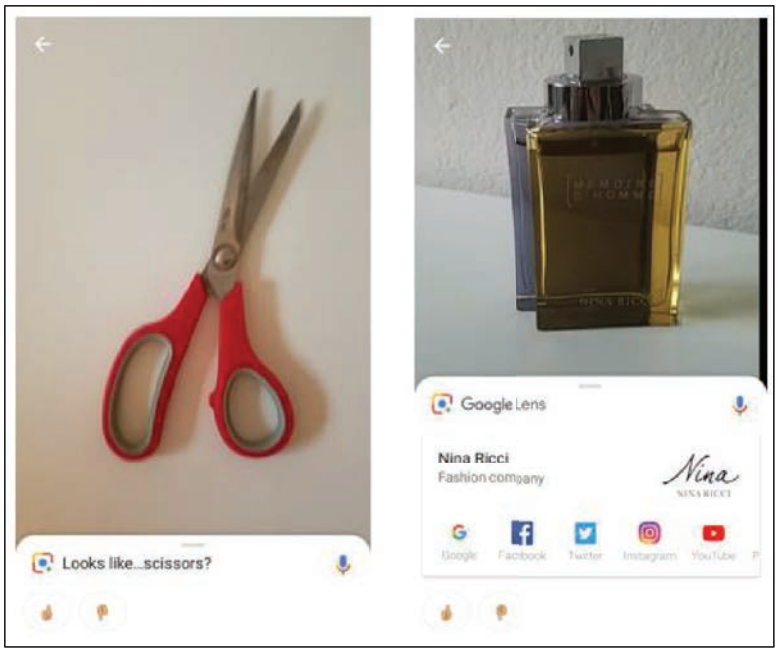


FIGURE 6.28 Two Examples of Using the Google Lens, a Service Based on Convolutional Deep Networks for Image Recognition. *Source:* ©2018 Google LLC, used with permission. Google and the Google logo are registered trademarks of Google LLC.

contact information from a business card image on the phone, identifying type of plants and breed of animals, identifying books and movies from their cover photos, and providing information (e.g., stores, theaters, shopping, reservations) about them. Figure 6.28 shows two examples of using the Google Lens app on an Android mobile device.

Even though later more accurate networks have been developed (e.g., He, Zhang, Ren, & Sun, 2015) in terms of efficiency and processing requirements (i.e., smaller number of layers and parameters), GoogLeNet is considered to be one of the best architectures to date. Apart from AlexNet and GoogLeNet, several other convolutional network architectures such as Residual Networks (ResNet), VGGNet, and Xception have been developed and contributed to the image-processing area, all relying on the ImageNet database.

In a May 2018 effort to address the labor-intensive task of labeling images on a large scale, Facebook published a *weakly supervised training* image recognition deep learning project (Mahajan et al., 2018). This project used hashtags made by the users on the images posted on Instagram as labels and trained a deep learning image recognition model based on that. The model was trained using 3.5 billion Instagram images labeled with around 17,000 hashtags using 336 GPUs working in parallel; the training procedure took a few weeks to be accomplished. A preliminary version of the model (trained using only 1 billion images and 1,500 hashtags) was then tested on the ImageNet benchmark data set and is reported to have outperformed the state-of-the-art models in terms of accuracy by more than 2 percent. This big achievement by Facebook surely will open doors to a new world of image processing using deep learning since it can dramatically increase the size of available image data sets that are labeled for training purposes.

Use of deep learning and advanced analytics methods to classify images has evolved into the recognition of human faces and has become a very popular application for a variety of purposes. It is discussed in Application Case 6.6.

Application Case 6.6 From Image Recognition to Face Recognition

Face recognition, although seemingly similar to image recognition, is a much more complicated undertaking. The goal of face recognition is to identify the individual as opposed to the class it belongs to (human), and this identification task needs to be performed on a nonstatic (i.e., moving person) 3D environment. Face recognition has been an active research field in AI for many decades with limited success until recently. Thanks to the new generation of algorithms (i.e., deep learning) coupled with large data sets and computational power, face recognition technology is starting to make a significant impact on real-world applications. From security to marketing, face recognition and the variety of applications/use cases of this technology are increasing at an astounding pace.

Some of the premier examples of face recognition (both in advancements in technology and in the creative use of the technology perspectives) come from China. Today in China, face recognition is a very hot topic both from business development and from application development perspectives. Face recognition has become a fruitful ecosystem with hundreds of start-ups in China. In personal and/or business settings, people in China are widely using and relying on devices whose security is based on automatic recognition of their faces.

As perhaps the largest scale practical application case of deep learning and face recognition in the world today, the Chinese government recently started a project known as "Sharp Eyes" that aims at establishing a nationwide surveillance system based on face recognition. The project plans to integrate security cameras already installed in public places with private cameras on buildings and to utilize AI and deep learning to analyze the videos from those cameras. With millions of cameras and billions of lines of code, China is building a high-tech authoritarian future. With this system, cameras in some cities can scan train and bus stations as well as airports to identify and catch China's most wanted suspected criminals. Billboard-size displays can show the faces of jaywalkers and list the names and pictures of people who do not pay their debts. Facial recognition scanners guard the entrances to housing complexes.

An interesting example of this surveillance system is the "shame game" (Mozur, 2018). An intersection south of Changhong Bridge in the city of Xiangyang previously was a nightmare. Cars drove fast, and jaywalkers darted into the street. Then, in the summer of 2017, the police put up cameras linked to facial recognition technology and a big outdoor screen. Photos of lawbreakers were displayed alongside their names and government identification numbers. People were initially excited to see their faces on the screen until propaganda outlets told them that this was a form of punishment. Using this, citizens not only became a subject of this shame game but also were assigned negative citizenship points. Conversely, on the positive side, if people are caught on camera showing good behavior, like picking up a piece of trash from the road and putting it into a trash can or helping an elderly person cross an intersection, they get positive citizenship points that can be used for a variety of small awards.

China already has an estimated 200 million surveillance cameras—four times as many as the United States. The system is mainly intended to be used for tracking suspects, spotting suspicious behavior, and predicting crimes. For instance, to find a criminal, the image of a suspect can be uploaded to the system, matching it against millions of faces recognized from videos of millions of active security cameras across the country. This can find individuals with a high degree of similarity. The system also is merged with a huge database of information on medical records, travel bookings, online purchases, and even social media activities of every citizen and can monitor practically everyone in the country (with 1.4 billion people), tracking where they are and what they are doing each moment (Denyer, 2018). Going beyond narrowly defined security purposes, the government expects Sharp Eyes to ultimately assign every individual in the country a "social credit score" that specifies to what extent she or he is trustworthy.

While such an unrestricted application of deep learning (i.e., spying on citizens) is against the privacy and ethical norms and regulations of many western countries, including the United States, it is becoming a common practice in countries with less restrictive privacy laws and concerns as in China. Even western countries have begun to plan on employing similar technologies in limited scales only for security and

crime prevention purposes. The FBI's Next Generation Identification System, for instance, is a lawful application of facial recognition and deep learning that compares images from crime scenes with a national database of mug shots to identify potential suspects.

Questions for Case 6.6

1. What are the technical challenges in face recognition?
2. Beyond security and surveillance purposes, where else do you think face recognition can be used?
3. What are the foreseeable social and cultural problems with developing and using face recognition technology?

Sources: Mozur, P. (2018, June 8). "Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras." *The New York Times*. **https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html**; Denyer, S. (2018, January). "Beijing Bets on Facial Recognition in a Big Drive for Total Surveillance." *The Washington Post*. **https://www.washingtonpost.com/news/world/wp/2018/01/07/feature/in-china-facial-recognition-is-sharp-end-of-a-drive-for-total-surveillance/?noredirect=on&utm_term=.e73091681b31**.

Text Processing Using Convolutional Networks

In addition to image processing, which was in fact the main reason for the popularity and development of convolutional networks, they have been shown to be useful in some large-scale text mining tasks as well. Especially since 2013, when Google published its **word2vec** project (Mikolov et al., 2013; Mikolov, Sutskever, Chen, Corrado, and Dean, 2013), the applications of deep learning for text mining have increased remarkably.

Word2vec is a two-layer neural network that gets a large text corpus as the input and converts each word in the corpus to a numeric vector of any given size (typically ranging from 100 to 1,000) with very interesting features. Although word2vec itself is not a deep learning algorithm, its outputs (word vectors also known as **word embeddings**) already have been widely used in many deep learning research and commercial projects as inputs.

One of the most interesting properties of word vectors created by the word2vec algorithm is maintaining the words' relative associations. For example, vector operations

$$vector(\text{'}King\text{'}) - vector(\text{'}Man\text{'}) + vector(\text{'}Woman\text{'})$$

and

$$vector(\text{'}London\text{'}) - vector(\text{'}England\text{'}) + vector(\text{'}France\text{'})$$

will result in a vector very close to *vector* (*'Queen'*) and *vector* (*'Paris'*), respectively. Figure 6.29 shows a simple vector representation of the first example in a two-dimensional vector space.

Moreover, the vectors are specified in such a way that those of a similar context are placed very close to each other in the *n*-dimensional vector space. For instance, in the word2vec model pretrained by Google using a corpus including about 100 billion words (taken from Google News), the closest vectors to the *vector* (*'Sweden'*) in terms of cosine distance, as shown in Table 6.2, identify European country names near the Scandinavian region, the same region in which Sweden is located.

Additionally, since word2vec takes into account the contexts in which a word has been used and the frequency of using it in each context in guessing the meaning of the word, it enables us to represent each term with its semantic context instead of just the syntactic/symbolic term itself. As a result, word2vec addresses several word variation issues that used to be problematic in traditional text mining activities. In other words,

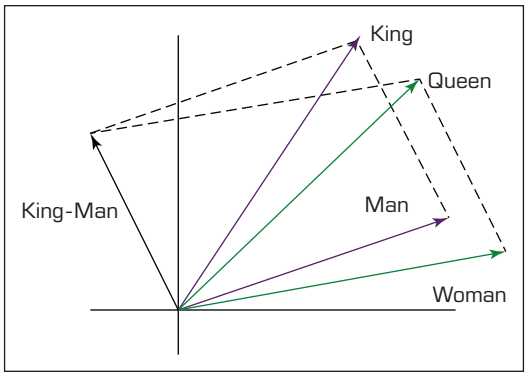


FIGURE 6.29 **Typical Vector Representation of Word Embeddings in a Two-Dimensional Space**

word2vec is able to handle and correctly represent words including typos, abbreviations, and informal conversations. For instance, the words *Frnce, Franse*, and *Frans* would all get roughly the same word embeddings as their original counterpart *France*. Word embeddings are also able to determine other interesting types of associations such as distinction of entities (e.g., *vector*['*human*'] − *vector*['*animal*']~*vector*['*ethics*']) or geopolitical associations (e.g., *vector*['*Iraq*'] − *vector*['*violence*']~*vector*['*Jordan*']).

By providing such a meaningful representation of textual data, in recent years, word2vec has driven many deep learning–based text mining projects in a wide range of contexts (e.g., medical, computer science, social media, marketing), and various types of deep networks have been applied to the word embeddings created by this algorithm to accomplish different objectives. Particularly, a large group of studies had developed convolutional networks applied to the word embeddings with the aim of *relation extraction* from textual data sets. Relation extraction is one of the subtasks of natural language processing (NLP) that focuses on determining whether two or more named entities recognized in the text form specific relationships (e.g., "A *causes* B"; "B *is caused by* A"). For instance, Zeng et al. (2014) developed a deep convolutional network (see Figure 6.30) to classify relations between specified entities in sentences. To this end, these researchers

TABLE 6.2 **Example of the word2vec Project Indicating the Closest Word Vectors to the Word "Sweden"**

Word	Cosine Distance
Norway	0.760124
Denmark	0.715460
Finland	0.620022
Switzerland	0.588132
Belgium	0.585635
Netherlands	0.574631
Iceland	0.562368
Estonia	0.547621
Slovenia	0.531408

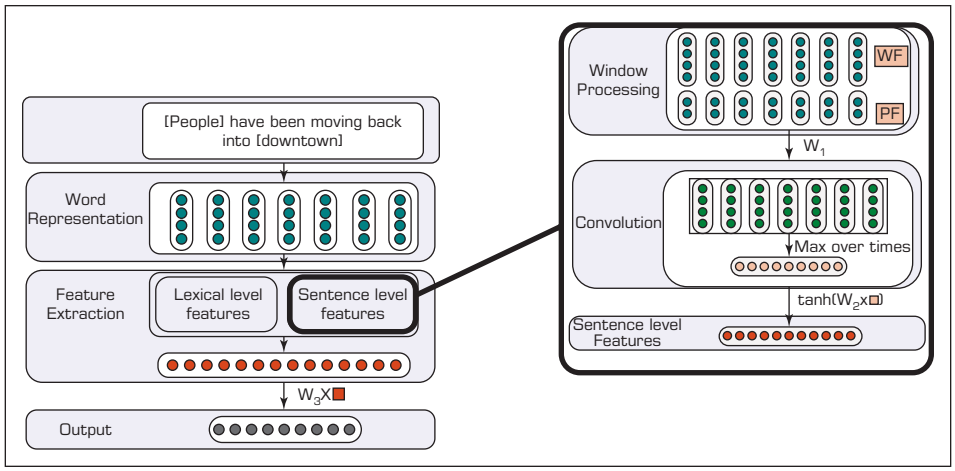


FIGURE 6.30 **CNN Architecture for Relation Extraction Task in Text Mining.**

used a matrix format to represent each sentence. Each column of the input matrices is in fact the word embedding (i.e., vector) associated with one of the words involved in the sentence. Zeng et al. then used a convolutional network, shown in the right box in Figure 6.30, to automatically learn the sentence-level features and concatenate those features (i.e., the output vector of the CNN) with some basic lexical features (e.g., the order of the two words of interest within the sentence and the left and right tokens for each of them). The concatenated feature vector then is fed into a classification layer with a *softmax* transfer function, which determines the type of relationship between the two words of interest among multiple predefined types. The softmax transfer function is the most common type of function to be used for classification layers, especially when the number of classes is more than two. For classification problems with only two outcome categories, log-sigmoid transfer functions are also very popular. The proposed approach by Zeng et al. was shown to correctly classify the relation between the marked terms in sentences of a sample data set with an 82.7 percent accuracy.

In a similar study, Nguyen and Grishman (2015) used a four-layer convolutional network with multiple kernel sizes in each convolution layer fed by the real-valued vectors of words included in sentences to classify the type of relationship between the two marked words in each sentence. In the input matrix, each row was the word embedding associated with a word in the same sequence in the sentence as the row number. In addition, these researchers included two more columns to the input matrices to represent the relative position of each word (either positive or negative) with regard to each of the marked terms. The automatically extracted features then were passed through a classification layer with softmax function for the type of relationship to be determined. Nguyen and Grishman trained their model using 8,000 annotated examples (with 19 predefined classes of relationships) and tested the trained model on a set of 2,717 validation data sets and achieved a classification accuracy of 61.32 percent (i.e., more than 11 times better performance than guessing).

Such text mining approaches using convolutional deep networks can be extended to various practical contexts. Again, the big challenge here, just as in image processing, is lack of sufficient large annotated data sets for supervised training of deep networks. A *distant supervision* method of training has been proposed (Mintz et al., 2009) to address this challenge. It suggests that large amounts of training data can be produced by aligning knowledge base (KB) facts with texts. In fact, this approach is based on the assumption that if a particular type of relation exists between an entity pair (e.g., "A" *is a component of* "B") in the KB, then every text document containing the mention of the

entity pair would express that relation. However, since this assumption was not very realistic, Riedel, Yao, and McCallum (2010) later relaxed it by modeling the problem as a *multi-instance learning* problem. They suggest assigning labels to a bag of instances rather than a single instance that can reduce the noise of the distant supervision method and create more realistic labeled training data sets (Kumar, 2017).

SECTION 6.7 REVIEW QUESTIONS

1. What is CNN?
2. For what type of applications can CNN be used?
3. What is convolution function in CNN and how does it work?
4. What is pooling in CNN? How does it work?
5. What is ImageNet and how does it relate to deep learning?
6. What is the significance of AlexNet? Draw and describe its architecture.
7. What is GoogLeNet? How does it work?
8. How does CNN process text? What are word embeddings, and how do they work?
9. What is word2vec, and what does it add to traditional text mining?

6.8 RECURRENT NETWORKS AND LONG SHORT-TERM MEMORY NETWORKS

Human thinking and understanding to a great extent relies on *context*. It is crucial for us, for example, to know that a particular speaker uses very sarcastic language (based on his previous speeches) to fully catch all the jokes that he makes. Or to understand the real meaning of the word *fall* (i.e., either *the season* or *to collapse*) in the sentence "It is a nice day of fall" without knowledge about the other words in the surrounding sentences would only be guessing, not necessarily understanding. Knowledge of context is typically formed based on observing events that happened in the past. In fact, human thoughts are persistent, and we use every piece of information we previously acquired about an event in the process of analyzing it rather than throwing away our past knowledge and thinking from scratch every time we face similar events or situations. Hence, there seems to be a recurrence in the way humans process information.

While deep MLP and convolutional networks are specialized for processing a *static* grid of values like an image or a matrix of word embeddings, sometimes the *sequence* of input values is also important to the operation of the network to accomplish a given task and hence should be taken into account. Another popular type of neural networks is **recurrent neural network (RNN)** (Rumelhart et al., 1986), which is specifically designed to process sequential inputs. An RNN basically models a *dynamic* system where (at least in one of its hidden neurons) the state of the system (i.e., output of a hidden neuron) at each time point t depends on both the inputs to the system at that time and its state at the previous time point $t - 1$. In other words, RNNs are the type of neural networks that have memory and that apply that memory to determine their future outputs. For instance, in designing a neural network to play chess, it is important to take into account several previous moves while training the network, because a wrong move by a player can lead to the eventual loss of the game in the subsequent 10–15 plays. Also, to understand the real meaning of a sentence in an essay, sometimes we need to rely on the information portrayed in the previous several sentences or paragraphs. That is, for a true understanding, we need the context built sequentially and collectively over time. Therefore, it is crucial to consider a memory element for the neural network that takes into account the effect of prior moves (in the chess example) and prior sentences and paragraphs (in the essay example) to determine the best output. This memory portrays and creates the context required for the learning and understanding.

In static networks like MLP-type CNNs, we are trying to find some functions (i.e., network weights and biases) that map the inputs to some outputs that are as close as possible to the actual target. In dynamic networks like RNNs, on the other hand, both inputs and outputs are sequences (patterns). Therefore, a dynamic network is a dynamic system rather than a function because its output depends not only on the input but also on the previous outputs. Most of the RNNs use the following general equation to define the values of their hidden units (Goodfellow et al., 2016).

$$a^{(t)} = f(a^{(t-1)}, p^{(t)}, \theta)$$

In this equation, $a^{(t)}$ represents the state of the system at time t, and $p^{(t)}$ and θ represent the input to the unit at time t and the parameters, respectively. Applying the same general equation for calculating the state of system at time $t - 1$, we will have:

$$a^{(t-1)} = f(a^{(t-2)}, p^{(t-1)}, \theta)$$

In other words:

$$a^{(t)} = f\left(f(a^{(t-2)}, p^{(t-1)}, \theta), p^{(t)}, \theta\right)$$

And this equation can be extended multiple times for any given sequence length. Graphically, a recurrent unit in a network can be depicted in a circuit diagram like the one shown in Figure 6.31. In this figure, D represents the *tap delay lines*, or simply the *delay* element of the network that, at each time point t, contains $a^{(t)}$, the previous output value of the unit. Sometimes instead of just one value, we store several previous output values in D to account for the effect of all of them. Also *iw* and *lw* represent the weight vectors applied to the input and the delay, respectively.

Technically speaking, any network with feedback can actually be called a *deep network*, because even with a single layer, the loop created by the feedback can be thought of as a static MLP-type network with many layers (see Figure 6.32 for a graphical illustration of this structure). However, in practice, each recurrent neural network would involve dozens of layers, each with feedback to itself, or even to the previous layers, which makes a recurrent neural network even deeper and more complicated.

Because of the feedbacks, computation of gradients in the recurrent neural networks would be somewhat different from the general backpropagation algorithm used

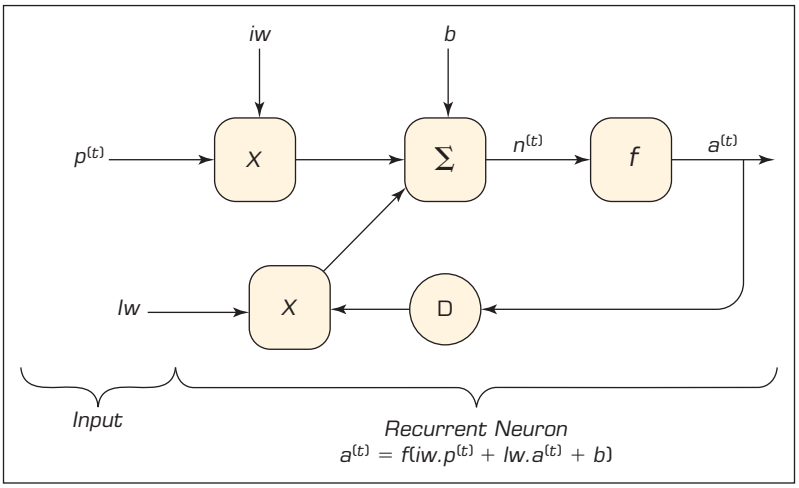


FIGURE 6.31 **Typical Recurrent Unit.**

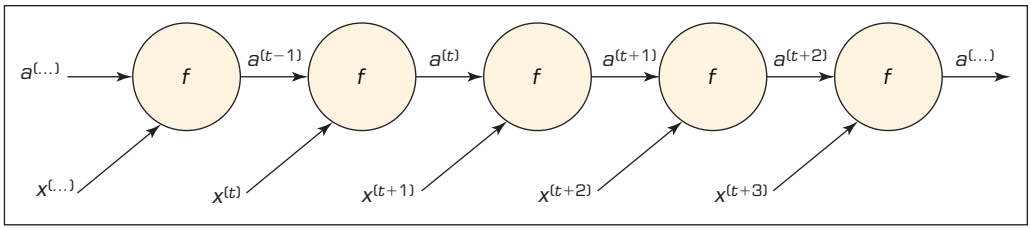


FIGURE 6.32 Unfolded View of a Typical Recurrent Network.

for the static MLP networks. There are two alternative approaches for computing the gradients in the RNNs, namely, real-time recurrent learning (RTRL) and backpropagation through time (BTT), whose explanation is beyond the scope of this chapter. Nevertheless, the general purpose remains the same; once the gradients have been computed, the same procedures are applied to optimize the learning of the network parameters.

The LSTM networks (Hochreiter & Schmidhuber, 1997) are variations of recurrent neural networks that today are known as the most effective sequence modeling technique and are the base of many practical applications. In a dynamic network, the weights are called the *long-term memory* while the feedbacks role is the *short-term memory*.

In essence, only the short-term memory (i.e., feedbacks; previous events) provides a network with the context. In a typical RNN, the information in the short-term memory is continuously replaced as new information is fed back into the network over time. That is why RNNs perform well when the gap between the relevant information and the place that is needed is small. For instance, for predicting the last word in the sentence "The referee blew his whistle," we just need to know a few words back (i.e., the referee) to correctly predict. Since in this case the gap between the relevant information (i.e., the referee) and where it is needed (i.e., to predict whistle) is small, an RNN network can easily perform this learning and prediction task.

However, sometimes the relevant information required to perform a task is far away from where it is needed (i.e., the gap is large). Therefore, it is quite likely that it would have already been replaced by other information in the short-term memory by the time it is needed for the creation of the proper context. For instance, to predict the last word in "I went to a carwash yesterday. It cost $5 to wash my *car*," there is a relatively larger gap between the relevant information (i.e., carwash) and where it is needed. Sometimes we may even need to refer to the previous paragraphs to reach the relevant information for predicting the true meaning of a word. In such cases, RNNs usually do not perform well since they cannot keep the information in their short-term memory for a long enough time. Fortunately, LSTM networks do not have such a shortcoming. The term *long short-term memory network* then refers to a network in which we are trying to remember what happened in the past (i.e., feedbacks; previous outputs of the layers) for a long enough time so that it can be used/leveraged in accomplishing the task when needed.

From an architectural viewpoint, the memory concept (i.e., remembering "what happened in the past") is incorporated in LSTM networks by incorporating four additional layers into the typical recurrent network architecture: three gate layers, namely *input gate, forget* (a.k.a. *feedback*) *gate*, and *output gate*, and an additional layer called **Constant Error Carousel (CEC)**, also known as *the state unit* that integrates those gates and interacts them with the other layers. Each gate is nothing but a layer with two inputs, one from the network input and the other a feedback from the final output of the whole network. The gates involve log-sigmoid transfer functions. Therefore, their outputs will be between 0 and 1 and describe how much of each component (either input, feedback, or output) should be let through the network. Also, CEC is a layer that falls between the

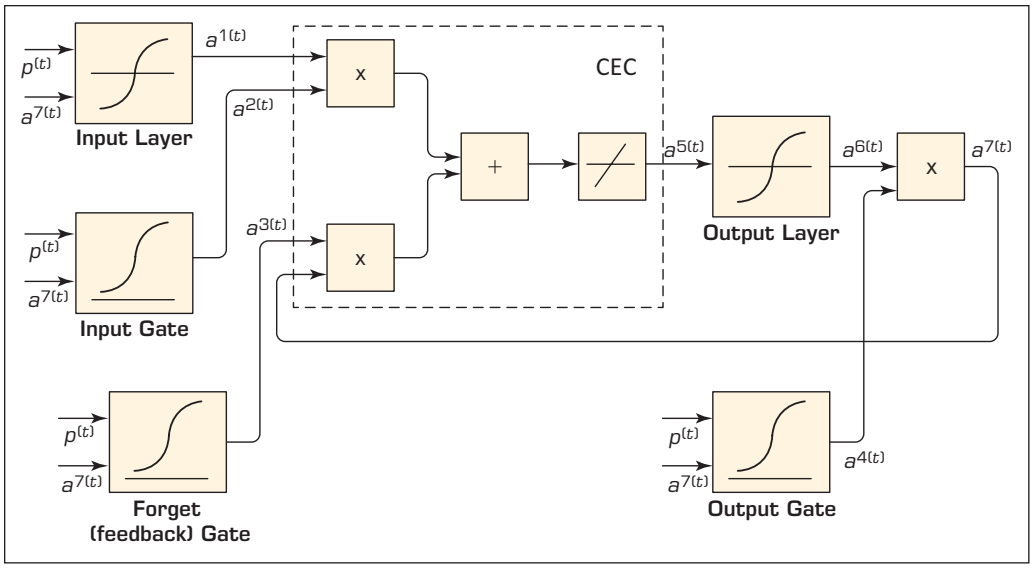


FIGURE 6.33 Typical Long Short-Term Memory (LSTM) Network Architecture.

input and the output layers in a recurrent network architecture and applies the gates outputs to make the short-term memory long.

To have a long short-term memory means that we want to keep the effect of previous outputs for a longer time. However, we typically do not want to indiscriminately remember everything that has happened in the past. Therefore, gating provides us with the capability of remembering prior outputs selectively. The input gate will allow selective inputs to the CEC; the forget gate will clear the CEC from the unwanted previous feedbacks; and the output gate will allow selective outputs from the CEC. Figure 6.33 shows a simple depiction of a typical LSTM architecture.

In summary, the gates in the LSTM are in charge of controlling the flow of information through the network and dynamically change the time scale of integration based on the input sequence. As a result, LSTM networks are able to learn long-term dependencies among the sequence of inputs more easily than the regular RNNs.

Application Case 6.7 illustrates the use of text processing in the context of understanding customer opinions and sentiments toward innovatively designing and developing new and improved products and services.

Application Case 6.7 Deliver Innovation by Understanding Customer Sentiments

Analyzing product and customer behavior provides valuable insights into what consumers want, how they interact with products, and where they encounter usability issues. These insights can lead to new feature designs and development or even new products.

Understanding customer sentiment and knowing what consumers truly think about products or a brand are traditional pain points. Customer journey analytics provides insights into these areas, yet these solutions are not all designed to integrate vital sources of unstructured data such as call center

(*Continued*)

Application Case 6.7 (Continued)

notes or social media feedback. In today's world, unstructured notes are part of core communications in virtually every industry, for example:

- Medical professionals record patient observations.
- Auto technicians write down safety information.
- Retailers track social media for consumer comments.
- Call centers monitor customer feedback and take notes.

Bringing together notes, which are usually available as free-form text, with other data for analysis has been difficult. That is because each industry has its own specific terms, slang, shorthand, and acronyms embedded in the data. Finding meaning and business insights first requires the text to be changed into a structured form. This manual process is expensive, time consuming, and prone to errors, especially as data scales to ever-increasing volumes. One way that companies can leverage notes without codifying the text is to use text clustering. This analytic technique quickly identifies common words or phrases for rapid insights.

Text and Notes Can Lead to New and Improved Products

Leveraging the insights and customer sentiment uncovered during a text and sentiment analysis can spark innovation. Companies such as vehicle manufacturers can use the intelligence to improve customer service and deliver an elevated customer experience. By learning what customers like and dislike about current products, companies can improve their design, such as adding new features to a vehicle to enhance the driving experience.

Forming word clusters also allows companies to identify safety issues. If an auto manufacturer sees that numerous customers are expressing negative sentiments about black smoke coming from their vehicle, the company can respond. Likewise, manufacturers can address safety issues that are a concern to customers. With comments grouped into buckets, companies have the ability to focus on specific customers who experienced a similar problem. This allows a company to, for instance, offer a rebate or special promotion to those who experienced black smoke.

Understanding sentiments can better inform a vehicle manufacturer's policies. For example, customers have different lifetime values. A customer who complains just once but has a very large lifetime value can be a more urgent candidate for complaint resolution than a customer with a lower lifetime value with multiple issues. One may have spent $5,000 buying the vehicle from a used vehicle lot. Another may have a history of buying new cars from the manufacturer and spent $30,000 to buy the vehicle on the showroom floor.

Analyzing Notes Enables High-Value Business Outcomes

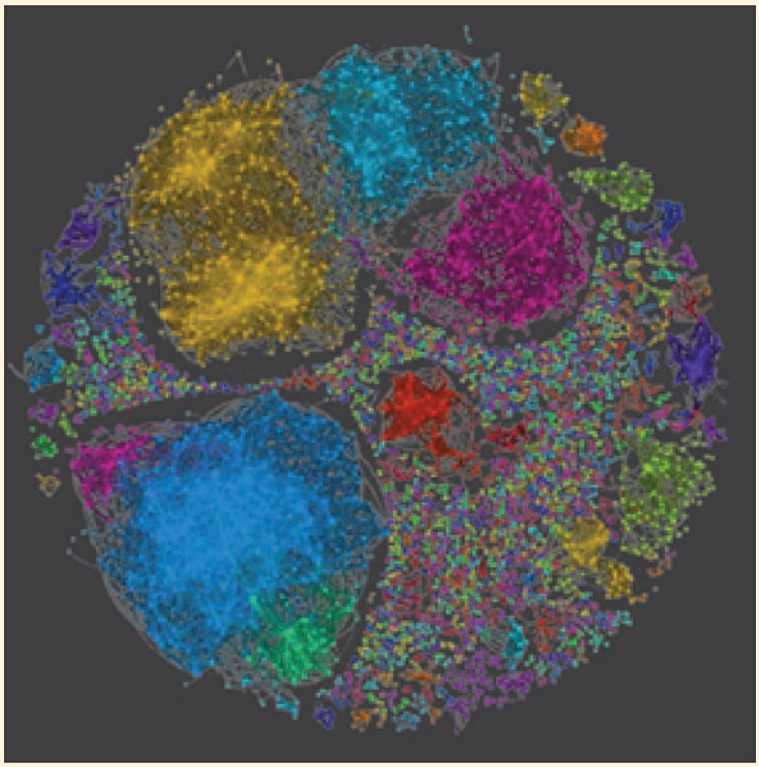
Managing the life cycle of products and services continues to be a struggle for most companies. The massive volumes of data now available have complicated life cycle management, creating new challenges for innovation. At the same time, the rapid rise of consumer feedback through social media has left businesses without a strategy for digesting, measuring, or incorporating the information into their product innovation cycle—meaning they miss a crucial amount of intelligence that reflects a customer's actual thoughts, feelings, and emotions.

Text and sentiment analysis is one solution to this problem. Deconstructing topics from masses of text allows companies to see what common issues, complaints, or positive or negative sentiments customers have about products. These insights can lead to high-value outcomes, such as improving products or creating new ones that deliver a better user experience, responding timely to safety issues, and identifying which product lines are most popular with consumers.

Example: Visualizing Auto Issues with "The Safety Cloud"

The Teradata Art of Analytics uses data science, Teradata® Aster® Analytics, and visualization techniques to turn data into one-of-a-kind artwork. To demonstrate the unique insights offered by text clustering, data scientists used the Art of Analytics to create "The Safety Cloud."

The scientists used advanced analytics algorithms on safety inspector and call center notes from an automobile manufacturer. The analytics identified and systematically extracted common words and phrases embedded in the data volumes. The blue cluster represents power steering failure. The pink is engine stalls. Yellow is black smoke in the exhaust.



Orange is brake failure. The manufacturer can use this information to gauge how big the problem is and whether it is safety related, and if so, then take actions to fix it.

For a visual summary, you can watch the video (**http://www.teradata.com/Resources/Videos/Art-of-Analytics-Safety-Cloud**).

Questions for Case 6.7

1. Why do you think sentiment analysis is gaining overwhelming popularity?
2. How does sentiment analysis work? What does it produce?
3. In addition to the specific examples in this case, can you think of other businesses and industries that can benefit from sentiment analysis? What is common among the companies that can benefit greatly from sentiment analysis?

Source: Teradata Case Study. "Deliver Innovation by Understanding Customer Sentiments." **http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB9859.pdf** (accessed August 2018). Used with permission.

LSTM Networks Applications

Since their emergence in the late 1990s (Hochreiter & Schmidhuber, 1997), LSTM networks have been widely used in many sequence modeling applications, including image captioning (i.e., automatically describing the content of images) (Vinyals, Toshev, Bengio, and Erhan, 2017, 2015; Xu et al., 2015), handwriting recognition and generation (Graves, 2013; Graves and Schmidhuber, 2009; Keysers et al. 2017), parsing (Liang et al. 2016; Vinyals, Kaiser, et al., 2015), speech recognition (Graves and Jaitly, 2014; Graves, Jaitly, and Mohamed, 2013; Graves, Mohamed, and Hinton, 2013), and machine translation (Bahdanau, Cho, and Bengio, 2014; Sutskever, Vinyals, and Le, 2014).

Currently, we are surrounded by multiple deep learning solutions working on the basis of speech recognition, such as Apple's Siri, Google Now, Microsoft's Cortana, and Amazon's Alexa, several of which we deal with on a daily basis (e.g., checking on the weather, asking for a Web search, calling a friend, and asking for directions on the map). Note taking is not a difficult, frustrating task anymore since we can easily record a speech or lecture, upload the digital recording on one of the several cloud-based speech-to-text service providers' platforms, and download the transcript in a few seconds. The Google cloud-based speech-to-text service, for example, supports 120 languages and their variants and has the ability to convert speech to text either in real time or using recorded audios. The Google service automatically handles the noise in the audio; accurately punctuates the transcripts with commas, question marks, and periods; and can be customized by the user to a specific context by getting a set of terms and phrases that are very likely to be used in a speech and recognizing them appropriately.

Machine translation refers to a subfield of AI that employs computer programs to translate speech or text from one language to another. One of the most comprehensive machine translation systems is the Google's Neural Machine Translation (GNMT) platform. GNMT is basically an LSTM network with eight encoder and eight decoder layers designed by a group of Google researchers in 2016 (Wu et al., 2016). GNMT is specialized for translating whole sentences at a time as opposed to the former version of Google Translate platform, which was a phrase-based translator. This network is capable of naturally handling the translation of rare words (that previously was a challenge in machine translation) by dividing the words into a set of common subword units. GNMT currently supports automatic sentence translations between more than 100 languages. Figure 6.34 shows how a sample sentence was translated from French to English by GNMT and a human translator. It also indicates how closely the GNMT translations between different language pairs were ranked by the human speakers compared with translations made by humans.

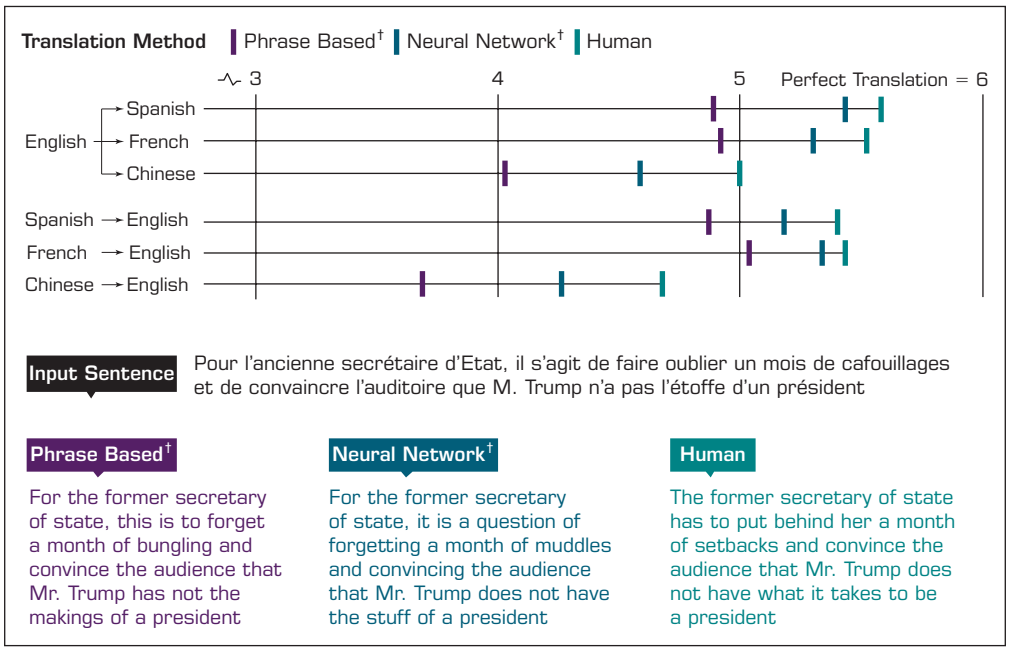


FIGURE 6.34 Example Indicating the Close-to-Human Performance of the Google Neural Machine Translator (GNMT)

Although machine translation has been revolutionized by the virtue of LSTMs, it encounters challenges that make it far from a fully automated high-quality translation. Like image-processing applications, there is a lack of sufficient training data (manually translated by humans) for many language pairs on which the network can be trained. As a result, translations between rare languages are usually done through a bridging language (mostly English) that may result in higher chances of error.

In 2014, Microsoft launched its Skype Translator service, a free voice translation service involving both speech recognition and machine translation with the ability of translating real-time conversations in 10 languages. Using this service, people speaking different languages can talk to each other in their own languages via a Skype voice or video call, and the system recognizes their voices and translates their every sentence through a translator bot in near real time for the other party. To provide more accurate translations, the deep networks used in the backend of this system were trained using conversational language (i.e., using materials such as translated Web pages, movie subtitles, and casual phrases taken from people's conversations in social networking Web sites) rather than the formal language commonly used in documents. The output of the speech recognition module of the system then goes through TrueText, a Microsoft technology for normalizing text that is capable of identifying mistakes and disfluencies (e.g., pauses during the speech or repeating some parts of speech, or adding fillers like "um" and "ah" when speaking) that people commonly conduct in their conversations and account for them for making better translations. Figure 6.35 shows the four-step process involved in the Skype Translator by Microsoft, each of which relies on the LSTM type of deep neural networks.

SECTION 6.8 REVIEW QUESTIONS

1. What is RNN? How does it differ from CNN?
2. What is the significance of "context," "sequence," and "memory" in RNN?
3. Draw and explain the functioning of a typical recurrent neural network unit.
4. What is the LSTM network, and how does it differ from RNNs?
5. List and briefly describe three different types of LSTM applications.
6. How do Google's Neural Machine Translation and Microsoft Skype Translator work?

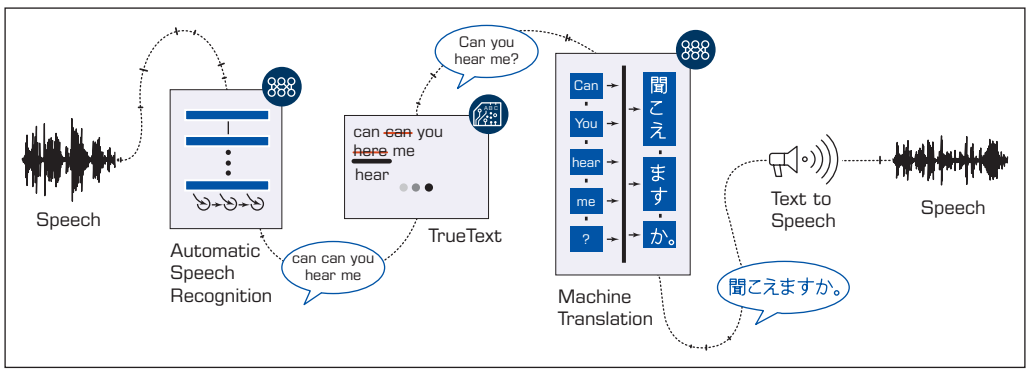


FIGURE 6.35 Four-Step Process of Translating Speech Using Deep Networks in the Microsoft Skype Translator.

6.9 COMPUTER FRAMEWORKS FOR IMPLEMENTATION OF DEEP LEARNING

Advances in deep learning owe its recent popularity, to a great extent, to advances in the software and hardware infrastructure required for its implementation. In the past few decades, GPUs have been revolutionized to support the playing of high-resolution videos as well as advanced video games and virtual reality applications. However, GPUs' huge processing potential had not been effectively utilized for purposes other than graphics processing up until a few years ago. Thanks to software libraries such as Theano (Bergstra et al., 2010), Torch (Collobert, Kavukcuoglu, and Farabet, 2011), Caffe (Jia et al., 2014), PyLearn2 (Goodfellow et al., 2013), Tensorflow (Abadi et al., 2016), and MXNet (Chen et al., 2015) developed with the purpose of programming GPUs for general-purpose processing (just as CPUs), and particularly for deep learning and analysis of Big Data, GPUs have become a critical enabler for the modern-day analytics. The operation of these libraries mostly relies on a parallel computing platform and application programming interface (API) developed by NVIDIA called *Compute Unified Device Architecture (CUDA)*, which enables software developers to use GPUs made by NVIDIA for general-purpose processing. In fact, each deep learning framework consists of a high-level scripting language (e.g., Python, R, Lua) and a library of deep learning routines usually written in C (for using CPUs) or CUDA (for using GPUs).

We next introduce some of the most popular software libraries used for deep learning by researchers and practitioners, including Torch, Caffe, Tensorflow, Theano, and Keras, and discuss some of their specific properties.

Torch

Torch (Collobert et al., 2011) is an open-source scientific computing framework (available at **www.torch.ch**) for implementing machine-learning algorithms using GPUs. The Torch framework is a library based on LuaJIT, a compiled version of the popular Lua programming language (**www.lua.org**). In fact, Torch adds a number of valuable features to Lua that make deep learning analyses possible; it enables supporting *n*-dimensional arrays (i.e., tensors), whereas tables (i.e., two-dimensional arrays) normally are the only data-structuring method used by Lua. Additionally, Torch includes routine libraries for manipulating (i.e., indexing, slicing, transposing) tensors, linear algebra, neural network functions, and optimization. More importantly, while Lua by default uses CPU to run the programs, Torch enables use of GPUs for running programs written in the Lua language.

The easy and extremely fast scripting properties of LuaJIT along with its flexibility have made Torch a very popular framework for practical deep learning applications such that today its latest version, Torch7, is widely used by a number of big companies in the deep learning area, including Facebook, Google, and IBM, in their research labs, as well as for their commercial applications.

Caffe

Caffe is another open-source deep learning framework (available at **http://caffe.berkeleyvision.org**) created by Yangqing Jia (2013), a PhD student at the University of California–Berkeley, which the Berkeley AI Research (BAIR) then further developed. Caffe has multiple options to be used as a high-level scripting language, including the command line, Python, and MATLAB interfaces. The deep learning libraries in Caffe are written in the C++ programming language.

In Caffe, everything is done using text files instead of code. That is, to implement a network, generally we need to prepare two text files with the *.prototxt* extension that are communicated by the Caffe engine via JavaScript Object Notation (JSON) format.

The first text file, known as the *architecture* file, defines the architecture of the network layer by layer, where each layer is defined by a name, a type (e.g., data, convolution, output), the names of its previous (bottom) and next (top) layers in the architecture, and some required parameters (e.g., kernel size and stride for a convolutional layer). The second text file, known as the *solver* file, specifies the properties of the training algorithm, including the learning rate, maximum number of iterations, and processing unit (CPU or GPU) to be used for training the network.

While Caffe supports multiple types of deep network architectures like CNN and LSTM, it is particularly known to be an efficient framework for image processing due to its incredible speed in processing image files. According to its developers, it is able to process over 60 million images per day (i.e., 1 ms/image) using a single NVIDIA K40 GPU. In 2017, Facebook released an improved version of Caffe called Caffe2 (**www.caffe2.ai**) with the aim of improving the original framework to be effectively used for deep learning architectures other than CNN and with a special emphasis on portability for performing cloud and mobile computations while maintaining scalability and performance.

TensorFlow

Another popular open-source deep learning framework is TensorFlow. It was originally developed and written in Python and C++ by the Google Brain Group in 2011 as *DistBelief*, but it was further developed into TensorFlow in 2015. TensorFlow at this time is the only deep learning framework that, in addition to CPUs and GPUs, supports Tensor Processing Units (TPUs), a type of processor developed by Google in 2016 for the specific purpose of neural network machine learning. In fact, TPUs were specifically designed by Google for the TensorFlow framework.

Although Google has not yet made TPUs available to the market, it is reported that it has used them in a number of its commercial services such as Google search, Street View, Google Photos, and Google Translate with significant improvements reported. A detailed study performed by Google shows that TPUs deliver 30 to 80 times higher performance per watt than contemporary CPUs and GPUs (Sato, Young, and Patterson, 2017). For example, it has been reported (Ung, 2016) that in Google Photos, an individual TPU can process over 100 million images per day (i.e., 0.86 ms/image). Such a unique feature will probably put TensorFlow way ahead of the other alternative frameworks in the near future as soon as Google makes TPUs commercially available.

Another interesting feature of TensorFlow is its visualization module, TensorBoard. Implementing a deep neural network is a complex and confusing task. TensorBoard refers to a Web application involving a handful of visualization tools to visualize network graphs and plot quantitative network metrics with the aim of helping users to better understand what is going on during training procedures and to debug possible issues.

Theano

In 2007, the Deep Learning Group at the University of Montreal developed the initial version of a Python library, Theano (**http://deeplearning.net/software/theano**), to define, optimize, and evaluate mathematical expressions involving multi-dimensional arrays (i.e., tensors) on CPU or GPU platforms. Theano was one of the first deep learning frameworks but later became a source of inspiration for the developers of TensorFlow. Theano and TensorFlow both pursue a similar procedure in the sense that in both a typical network implementation involves two sections: in the first section, a computational graph is built by defining the network variables and operations to be done on them; and the second section runs that graph (in Theano by compiling the graph into a *function* and in TensorFlow by creating a *session*). In fact, what happens in these libraries is that the user defines the structure of the network by providing some simple and symbolic

syntax understandable even for beginners in programming, and the library automatically generates appropriate codes in either C (for processing on CPU) or CUDA (for processing on GPU) to implement the defined network. Hence, users without any knowledge of programming in C or CUDA and with just a minimum knowledge of Python are able to efficiently design and implement deep learning networks on the GPU platforms.

Theano also includes some built-in functions to visualize computational graphs as well as to plot the network performance metrics even though its visualization features are not comparable to TensorBoard.

Keras: An Application Programming Interface

While all described deep learning frameworks require users to be familiar with their own syntax (through reading their documentations) to be able to successfully train a network, fortunately there are some easier, more user-friendly ways to do so. **Keras** (**https://keras.io/**) is an open-source neural network library written in Python that functions as a high-level application programming interface (API) and is able to run on top of various deep learning frameworks including Theano and TensorFlow. In essence, Keras just by getting the key properties of network building blocks (i.e., type of layers, transfer functions, and optimizers) via an extremely simple syntax automatically generates syntax in one of the deep learning frameworks and runs that framework in the backend. While Keras is efficient enough to build and run general deep learning models in just a few minutes, it does not provide several advanced operations provided by TensorFlow or Theano. Therefore, in dealing with special deep network models that require advanced settings, one still needs to directly use those frameworks instead of Keras (or other APIs such as *Lasagne*) as a proxy.

SECTION 6.9 REVIEW QUESTIONS

1. Despite the short tenure of deep learning implementation, why do you think there are several different computing frameworks for it?
2. Define *CPU, NVIDIA*, CUDA, and *deep learning*, and comment on the relationship between them.
3. List and briefly define the characteristics of different deep learning frameworks.
4. What is Keras, and how is it different from the other frameworks?

6.10 COGNITIVE COMPUTING

We are witnessing a significant increase in the way technology is evolving. Things that once took decades are now taking months, and the things that we see only in SciFi movies are becoming reality, one after another. Therefore, it is safe to say that in the next decade or two, technological advancements will transform how people live, learn, and work in a rather dramatic fashion. The interactions between humans and technology will become intuitive, seamless, and perhaps transparent. Cognitive computing will have a significant role to play in this transformation. Generally speaking, cognitive computing refers to the computing systems that use mathematical models to emulate (or partially simulate) the human cognition process to find solutions to complex problems and situations where the potential answers can be imprecise. While the term *cognitive computing* is often used interchangeably with AI and smart search engines, the phrase itself is closely associated with IBM's cognitive computer system Watson and its success on the television show *Jeopardy!* Details on Watson's success on *Jeopardy!* can be found in Application Case 6.8.

According to Cognitive Computing Consortium (2018), cognitive computing makes a new class of problems computable. It addresses highly complex situations that are

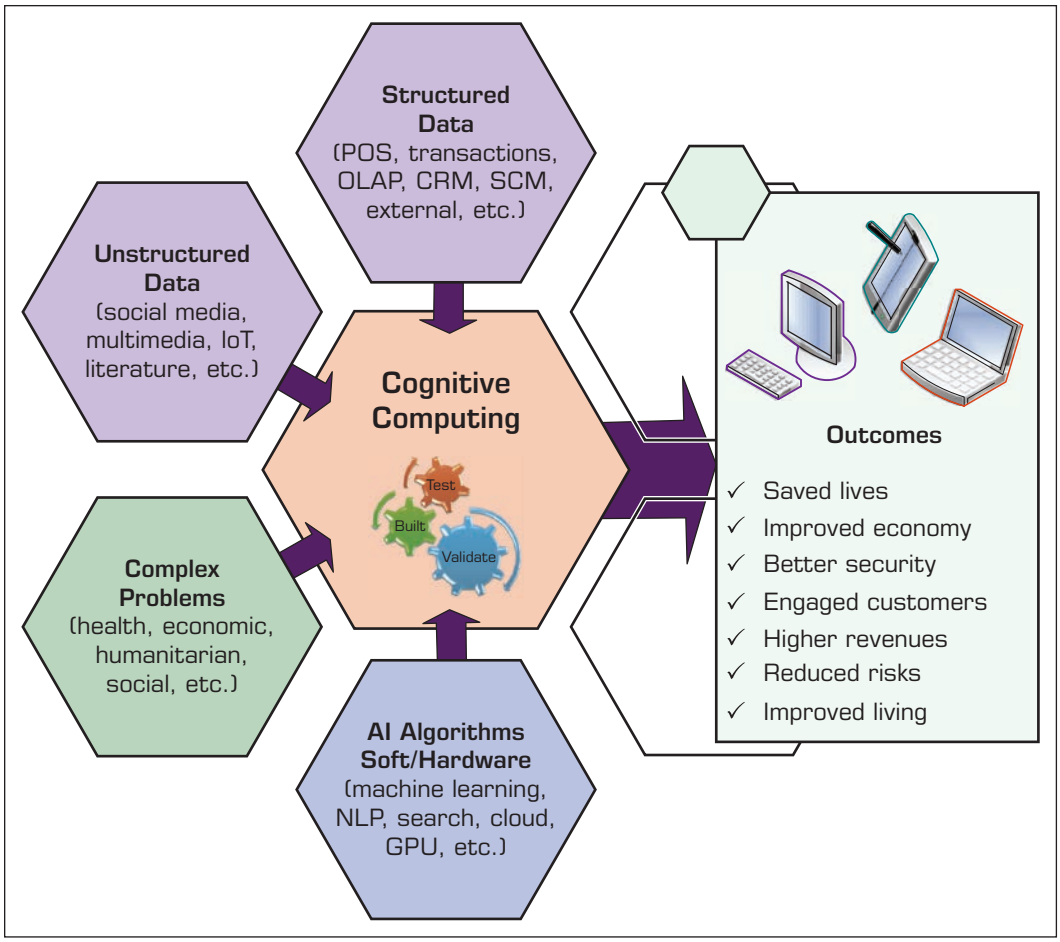


FIGURE 6.36 **Conceptual Framework for Cognitive Computing and Its Promises.**

characterized by ambiguity and uncertainty; in other words, it handles the kinds of problems that are thought to be solvable by human ingenuity and creativity. In today's dynamic, information-rich, and unstable situations, data tend to change frequently, and they often conflict. The goals of users evolve as they learn more and redefine their objectives. To respond to the fluid nature of users' understanding of their problems, the cognitive computing system offers a synthesis not just of information sources but also of influences, contexts, and insights. To achieve such a high-level of performance, cognitive systems often need to weigh conflicting evidence and suggest an answer that is "best" rather than "right." Figure 6.36 illustrates a general framework for cognitive computing where data and AI technologies are used to solve complex real-world problems.

How Does Cognitive Computing Work?

As one would guess from the name, cognitive computing works much like a human thought process, reasoning mechanism, and cognitive system. These cutting-edge computation systems can find and synthesize data from various information sources and weigh context and conflicting evidence inherent in the data to provide the best possible answers to a given question or problem. To achieve this, cognitive systems include self-learning technologies that use data mining, pattern recognition, deep learning, and NLP to mimic the way the human brain works.

Using computer systems to solve the types of problems that humans are typically tasked with requires vast amounts of structured and unstructured data fed to machine-learning algorithms. Over time, cognitive systems are able to refine the way in which they learn and recognize patterns and the way they process data to become capable of anticipating new problems and modeling and proposing possible solutions.

To achieve those capabilities, cognitive computing systems must have the following key attributes as defined by the Cognitive Computing Consortium (2018):

- **Adaptive:** Cognitive systems must be flexible enough to learn as information changes and goals evolve. The systems must be able to digest dynamic data in real time and make adjustments as the data and environment change.
- **Interactive:** Human-computer interaction (HCI) is a critical component in cognitive systems. Users must be able to interact with cognitive machines and define their needs as those needs change. The technologies must also be able to interact with other processors, devices, and cloud platforms.
- **Iterative and stateful:** Cognitive computing technologies can also identify problems by asking questions or pulling in additional data if a stated problem is vague or incomplete. The systems do this by maintaining information about similar situations that have previously occurred.
- **Contextual:** Understanding context is critical in thought processes, so cognitive systems must understand, identify, and mine contextual data, such as syntax, time, location, domain, requirements, and a specific user's profile, tasks, or goals. Cognitive systems may draw on multiple sources of information, including structured and unstructured data and visual, auditory, or sensor data.

How Does Cognitive Computing Differ from AI?

Cognitive computing is often used interchangeably with AI, the umbrella term used for technologies that rely on data and scientific methods/computations to make (or help/support in making) decisions. But there are differences between the two terms, which can largely be found within their purposes and applications. AI technologies include—but are not limited to—machine learning, neural computing, NLP, and, most recently, deep learning. With AI systems, especially in machine-learning systems, data are fed into the algorithm for processing (an iterative and time-demanding process that is often called *training*) so that the systems "learn" variables and interrelationships among those variables so that it can produce predictions (or characterizations) about a given complex problem or situation. Applications based on AI and cognitive computing include intelligent assistants, such as Amazon's Alexa, Google Home, and Apple's Siri. A simple comparison between cognitive computing and AI is given in Table 6.3 (Reynolds and Feldman, 2014; CCC, 2018).

As can be seen in Table 6.3, the differences between AI and cognitive computing are rather marginal. This is expected because cognitive computing is often characterized as a subcomponent of AI or an application of AI technologies tailored for a specific purpose. AI and cognitive computing both utilize similar technologies and are applied to similar industry segments and verticals. The main difference between the two is the purpose: while cognitive computing is aimed at helping humans to solve complex problems, AI is aimed at automating processes that are performed by humans; at the extreme, AI is striving to replace humans with machines for tasks requiring "intelligence," one at a time.

In recent years, cognitive computing typically has been used to describe AI systems that aim to simulate human thought process. Human cognition involves real-time analysis of environment, context, and intent among many other variables that inform a person's ability to solve problems. A number of AI technologies are required for a computer system to build cognitive models that mimic human thought processes, including machine learning, deep learning, neural networks, NLP, text mining, and sentiment analysis.

TABLE 6.3 Cognitive Computing versus Artificial Intelligence (AI)

Characteristic	Cognitive Computing	Artificial Intelligence (AI)
Technologies used	• Machine learning • Natural language processing • Neural networks • Deep learning • Text mining • Sentiment analysis	• Machine learning • Natural language processing • Neural networks • Deep learning
Capabilities offered	Simulate human thought processes to assist humans in finding solutions to complex problems	Find hidden patterns in a variety of data sources to identify problems and provide potential solutions
Purpose	Augment human capability	Automate complex processes by acting like a human in certain situations
Industries	Customer service, marketing, healthcare, entertainment, service sector	Manufacturing, finance, healthcare, banking, securities, retail, government

In general, cognitive computing is used to assist humans in their decision-making process. Some examples of cognitive computing applications include supporting medical doctors in their treatment of disease. IBM Watson for Oncology, for example, has been used at Memorial Sloan Kettering Cancer Center to provide oncologists evidence-based treatment options for cancer patients. When medical staff input questions, Watson generates a list of hypotheses and offers treatment options for doctors to consider. Whereas AI relies on algorithms to solve a problem or to identify patterns hidden in data, cognitive computing systems have the loftier goal of creating algorithms that mimic the human brain's reasoning process to help humans solve an array of problems as the data and the problems constantly change.

In dealing with complex situations, context is important, and cognitive computing systems make context computable. They identify and extract context features such as time, location, task, history, or profile to present a specific set of information that is appropriate for an individual or for a dependent application engaged in a specific process at a specific time and place. According to the Cognitive Computing Consortium, they provide machine-aided serendipity by wading through massive collections of diverse information to find patterns and then apply those patterns to respond to the needs of the user at a particular moment. In a sense, cognitive computing systems aim at redefining the nature of the relationship between people and their increasingly pervasive digital environment. They may play the role of assistant or coach for the user, and they may act virtually autonomously in many problem-solving situations. The boundaries of the processes and domains these systems can affect are still elastic and emergent. Their output may be prescriptive, suggestive, instructive, or simply entertaining.

In the short time of its existence, cognitive computing has proved to be useful in many domain and complex situations and is evolving into many more. The typical use cases for cognitive computing include the following:

- Development of smart and adaptive search engines
- Effective use of natural language processing
- Speech recognition
- Language translation
- Context-based sentiment analysis

- Face recognition and facial emotion detection
- Risk assessment and mitigation
- Fraud detection and mitigation
- Behavioral assessment and recommendations

Cognitive analytics is a term that refers to cognitive computing–branded technology platforms, such as IBM Watson, that specialize in processing and analyzing large, unstructured data sets. Typically, word processing documents, e-mails, videos, images, audio files, presentations, Web pages, social media, and many other data formats need to be manually tagged with metadata before they can be fed into a traditional analytics engine and Big Data tools for computational analyses and insight generation. The principal benefit of utilizing cognitive analytics over those traditional Big Data analytics tools is that for cognitive analytics such data sets do not need to be pretagged. Cognitive analytics systems can use machine learning to adapt to different contexts with minimal human supervision. These systems can be equipped with a chatbot or search assistant that understands queries, explains data insights, and interacts with humans in human languages.

Cognitive Search

Cognitive search is the new generation search method that uses AI (advanced indexing, NLP, and machine learning) to return results that are much more relevant to users. Forrester defines cognitive search and knowledge discovery solutions as "a new generation of enterprise search solutions that employ AI technologies such as natural language processing and machine learning to ingest, understand, organize, and query digital content from multiple data sources" (Gualtieri, 2017). Cognitive search creates searchable information out of nonsearchable content by leveraging cognitive computing algorithms to create an indexing platform.

Searching for information is a tedious task. Although current search engines do a very good job in finding relevant information in a timely manner, their sources are limited to publically available data over the Internet. Cognitive search proposes the next generation of search tailored for use in enterprises. It is different from traditional search because, according to Gualtieri (2017), it:

- **Can handle a variety of data types.** Search is no longer just about unstructured text contained in documents and in Web pages. Cognitive search solutions can also accommodate structured data contained in databases and even nontraditional enterprise data such as images, video, audio, and machine-/sensor-generated logs from IoT devices.
- **Can contextualize the search space.** In information retrieval, the context is important. Context takes the traditional syntax-/symbol-driven search to a new level where it is defined by semantics and meaning.
- **Employ advanced AI technologies.** The distinguishing characteristic of cognitive search solutions is that they use NLP and machine learning to understand and organize data, predict the intent of the search query, improve the relevancy of results, and automatically tune the relevancy of results over time.
- **Enable developers to build enterprise-specific search applications.** Search is not just about a text box on an enterprise portal. Enterprises build search applications that embed search in customer 360 applications, pharma research tools, and many other business process applications. Virtual digital assistants such as Amazon Alexa, Google Now, and Siri would be useless without powerful searches behind the scenes. Enterprises wishing to build similar applications for their customers will also benefit from cognitive search solutions. Cognitive search solutions provide software development kits (SDKs), APIs, and/or visual design tools that allow developers to embed the power of the search engine in other applications.

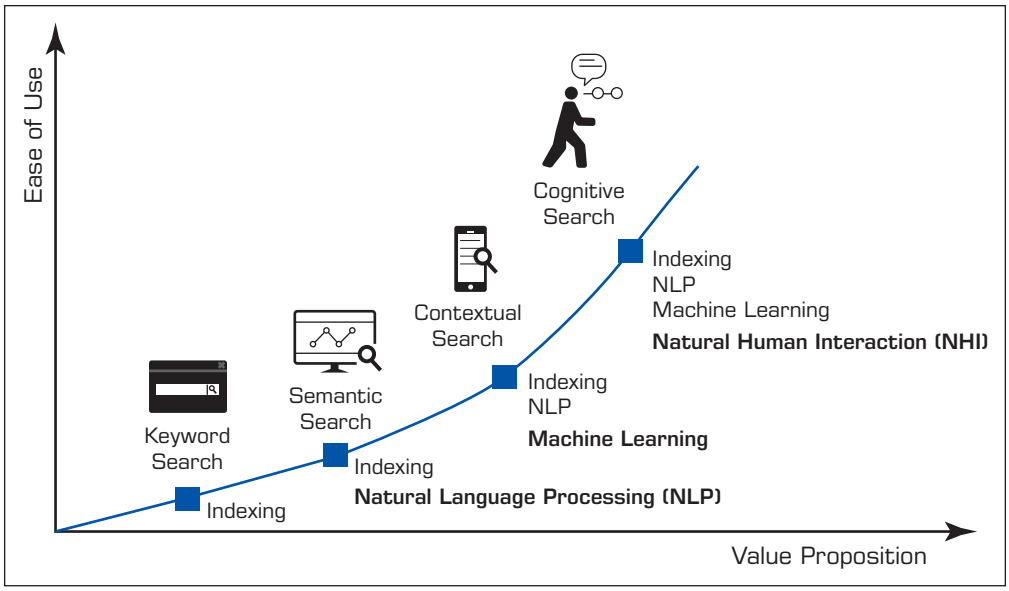


FIGURE 6.37 **Progressive Evolution of Search Methods.**

Figure 6.37 shows the progressive evolution of search methods from good old keyword search to modern-day cognitive search on two dimensions—ease of use and value proposition.

IBM Watson: Analytics at Its Best

IBM Watson is perhaps the smartest computer system built to date. Since the emergence of computers and subsequently AI in the late 1940s, scientists have compared the performance of these "smart" machines with human minds. Accordingly, in the mid- to late-1990s, IBM researchers built a smart machine and used the game of chess (generally credited as the game of smart humans) to test its ability against the best of human players. On May 11, 1997, an IBM computer called Deep Blue beat the world chess grandmaster after a six-game match series: two wins for Deep Blue, one for the champion, and three draws. The match lasted several days and received massive media coverage around the world. It was the classic plot line of human versus machine. Beyond the chess contest, the intention of developing this kind of computer intelligence was to make computers able to handle the kinds of complex calculations needed to help discover new drugs and to do the broad financial modeling needed to identify trends and do risk analysis, handle large database searches, and perform massive calculations needed in advanced fields of science.

After a couple of decades, IBM researchers came up with another idea that was perhaps more challenging: a machine that could not only play the American TV quiz show *Jeopardy!* but also beat the best of the best. Compared to chess, *Jeopardy!* is much more challenging. While chess is well structured and has very simple rules and therefore is a very good match for computer processing, *Jeopardy!* is neither simple nor structured. *Jeopardy!* is a game designed to test human intelligence and creativity. Therefore, a computer designed to play the game needed to be a cognitive computing system that can work and think like a human. Making sense of imprecision inherent in human language was the key to success.

In 2010, an IBM research team developed Watson, an extraordinary computer system—a novel combination of advanced hardware and software—designed to answer questions posed in natural human language. The team built Watson as part of the DeepQA project and named it after IBM's first president, Thomas J. Watson. The team that built Watson was looking for a major research challenge: one that could rival the scientific and popular interest of Deep Blue and would have clear relevance to IBM's business interests. The goal was to advance computational science by exploring new ways for computer technology to affect science, business, and society at large. Accordingly, IBM research undertook a challenge to build Watson as a computer system that could compete at the human champion level in real time on *Jeopardy!* The team wanted to create a real-time automatic contestant on the show capable of listening, understanding, and responding, not merely a laboratory exercise. Application Case 6.8 provides some of the details on IBM Watson's participation in the game show.

Application Case 6.8 IBM Watson Competes against the Best at *Jeopardy!*

In 2011, to test its cognitive abilities, Watson competed on the quiz show *Jeopardy!* in the first-ever human-versus-machine matchup for the show. In a two-game, combined-point match (broadcast in three *Jeopardy!* episodes during February 14–16), Watson beat Brad Rutter, the highest all-time money winner on *Jeopardy!* and Ken Jennings, the record holder for the longest championship streak (75 days). In these episodes, Watson consistently outperformed its human opponents on the game's signaling device, but it had trouble responding to a few categories, notably those having short clues containing only a few words. Watson had access to 200 million pages of structured and unstructured content, consuming four terabytes of disk storage. During the game, Watson was not connected to the Internet.

Meeting the *Jeopardy!* challenge required advancing and incorporating a variety of text mining and NLP technologies, including parsing, question classification, question decomposition, automatic source acquisition and evaluation, entity and relationship detection, logical form generation, and knowledge representation and reasoning. Winning at *Jeopardy!* required accurately computing confidence in answers. The questions and content are ambiguous and noisy, and none of the individual algorithms is perfect. Therefore, each component must produce a confidence in its output, and individual component confidences must be combined to compute the overall confidence of the final answer. The final confidence is used to determine whether the computer system should risk choosing to answer at all. In *Jeopardy!* this confidence is used to determine whether the computer will "ring in" or "buzz in" for a question. The confidence must be computed during the time the question is read and before the opportunity to buzz in. This is roughly between one and six seconds with an average around three seconds.

Watson was an excellent example for the rapid advancement of the computing technology and what it is capable of doing. Although still not as creatively/natively smart as human beings, computer systems like Watson are evolving to change the world we are living in, hopefully for the better.

Questions for Case 6.8

1. In your opinion, what are the most unique features about Watson?
2. In what other challenging games would you like to see Watson compete against humans? Why?
3. What are the similarities and differences between Watson's and humans' intelligence?

Sources: Ferrucci, D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, D. Kalyanpur, A. Lally, J. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. (2010). "Building Watson: An Overview of the DeepQA Project." *AI Magazine, 31*(3), pp. 59–79; IBM Corporation. (2011). "The DeepQA Project." **https://researcher.watson.ibm.com/researcher/view_group.php?id=2099** (accessed May 2018).

How Does Watson Do It?

What is under the hood of Watson? How does it do what it does? The system behind Watson, which is called DeepQA, is a massively parallel, text mining–focused, probabilistic evidence–based computational architecture. For the *Jeopardy!* challenge, Watson used more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique the IBM team used was how it combined them in DeepQA such that overlapping approaches could bring their strengths to bear and contribute to improvements in accuracy, confidence, and speed.

DeepQA is architecture with an accompanying methodology that is not specific to the *Jeopardy!* challenge. These are the overarching principles in DeepQA:

- **Massive parallelism.** Watson needed to exploit massive parallelism in the consideration of multiple interpretations and hypotheses.
- **Many experts.** Watson needed to be able to integrate, apply, and contextually evaluate a wide range of loosely coupled probabilistic questions and content analytics.
- **Pervasive confidence estimation.** No component of Watson committed to an answer; all components produced features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learned how to stack and combine the scores.
- **Integration of shallow and deep knowledge.** Watson needed to balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

Figure 6.38 illustrates the DeepQA architecture at a very high level. More technical details about the various architectural components and their specific roles and capabilities can be found in Ferrucci et al. (2010).

What Is the Future for Watson?

The *Jeopardy!* challenge helped IBM address requirements that led to the design of the DeepQA architecture and the implementation of Watson. After three years of intense research and development by a core team of about 20 researchers, as well as a significant

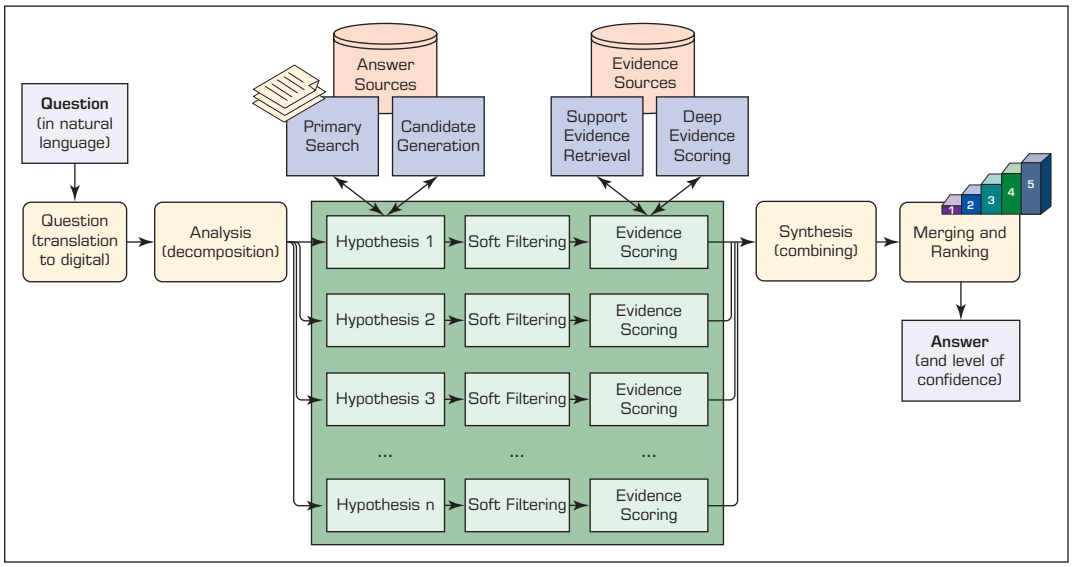


FIGURE 6.38 A High-Level Depiction of DeepQA Architecture

R&D budget, Watson managed to perform at human expert levels in terms of precision, confidence, and speed on the *Jeopardy!* quiz show.

After the show, the big question was "So what now?" Was developing Watson all for a quiz show? Absolutely not! Showing the rest of the world what Watson (and the cognitive system behind it) could do became an inspiration for the next generation of intelligent information systems. For IBM, it was a demonstration of what is possible with cutting-edge analytics and computational sciences. The message is clear: If a smart machine can beat the best of the best in humans at what they are the best at, think about what it can do for your organizational problems.

The innovative and futuristic technologies that made Watson one of the most acclaimed technological advances of this decade are being leveraged as computational foundation for several tools to analyze and characterize unstructured data for prediction-type problems. These experimental tools include Tone Analyzer and Personality Insights. Using textual content, these tools have shown the ability to predict outcomes of complex social events and globally popular competitions.

WATSON PREDICTS THE WINNER OF 2017 EUROVISION SONG CONTEST. A tool developed on the foundations of IBM Watson, Watson Tone Analyzer, uses computational linguistics to identify *tone* in written text. Its broader goal is to have business managers use the Tone Analyzer to understand posts, conversations, and communications of target customer populations and to respond to their needs and wants in a timely manner. One could, for example, use this tool to monitor social media and other Web-based content, including wall posts, tweets, product reviews, and discussion boards as well as longer documents such as articles and blog posts. Or one could use it to monitor customer service interactions and support related conversations. Although it sounds as if any other text-based detection system can build on sentiment analysis, Tone Analyzer differs from these systems in that it analyzes and characterizes textual content. Watson Tone Analyzer measures social tendencies and opinions, using a version of the Big-5, the five categories of personality traits (i.e., openness, agreeableness, conscientiousness, extroversion, and neuroticism), along with other emotional categories to detect the tone in a given textual content. As an example, Slowey (2017b) used IBM's Watson Tone Analyzer to predict the winner of the 2017 Eurovision Songs Contest. Using nothing but the lyrics of the previous years' competitions, Slowey discovered a pattern that suggested most winners had high levels of agreeableness and conscientiousness. The results (produced before the contest) indicated that Portugal would win the contest, and that is exactly what happened. Try it out yourself:

- Go to Watson Tone Analyzer (**https://tone-analyzer-demo.ng.bluemix.net**).
- Copy and paste your own text in the provided text entry field.
- Click "Analyze."
- Observe the summary results as well as the specific sentences where specific tones are the strongest

Another tool built on the linguistic foundations of IBM Watson is Watson Personality Insight, which seems to work quite similar to Watson Tone Analyzer. In another fun application case, Slowey (2017a) used Watson Personality Insight to predict the winner of the best picture category at the 2017 Oscar Academy Awards. Using the scripts of the movies from the past years, Slowey developed a generalized profile for winners and then compared that profile to those of the newly nominated movies to identify the upcoming winner. Although in this case, Slowey incorrectly predicted *Hidden Figures* as the winner, the methodology she followed was unique and innovative and hence deserves credit. To try Watson Personality Insight tool yourself, just go to **https://personality-insights-demo.ng.bluemix.net/**, copy and paste your own textual content into the "Body of Text" section, and observe the outcome.

One of the worthiest endeavors for Watson (or Watson-like large-scale cognitive computing systems) is to help doctors and other medical professionals to diagnose diseases and identify the best treatment options that would work for an individual patient. Although Watson is new, this very novel and worthy task is not new to the world of computing. In the early 1970s, several researchers at Stanford University developed a computer system, MYCIN, to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics with the dosage adjusted for the specifics of an individual patient (Buchanan and Shortliffe, 1984). This six-year effort relied on a rule-based expert system, a type of AI system, where the diagnoses and treatment knowledge nuggets/rules were elicited from a large number of experts (i.e., doctors with ample experience in the specific medical domain). The resulting system was then tested on new patients, and its performance was compared to those of the experienced doctors used as the knowledge sources/experts. The results favored MYCIN, providing a clear indication that properly designed and implemented AI-based computer systems can meet and often exceed the effectiveness and efficiency of even the best medical experts. After more than four decades, Watson is now trying to pick up where MYCIN left the mission of using smart computer systems to improve the health and well-being of humans by helping doctors with the contextual information that they need to better and more quickly diagnose and treat their patients.

The first industry targeted to utilize Watson was healthcare, followed by security, finance, retail, education, public services, and research. The following sections provide short descriptions of what Watson can do (and, in many cases, is doing) for these industries.

HEALTHCARE AND MEDICINE The challenges that healthcare is facing today are rather big and multifaceted. With the aging U.S. population, which may be partially attributed to better living conditions and advanced medical discoveries fueled by a variety of technological innovations, demand for healthcare services is increasing faster than the supply of resources. As we all know, when there is an imbalance between demand and supply, prices go up and quality suffers. Therefore, we need cognitive systems like Watson to help decision makers optimize the use of their resources in both clinical and managerial settings.

According to healthcare experts, only 20 percent of the knowledge that physicians use to diagnose and treat patients is evidence based. Considering that the amount of medical information available is doubling every five years and that much of these data are unstructured, physicians simply do not have time to read every journal that can help them keep up-to-date with the latest advances. Given the growing demand for services and the complexity of medical decision making, how can healthcare providers address these problems? The answer could be to use Watson or similar cognitive systems that have the ability to help physicians in diagnosing and treating patients by analyzing large amounts of data—both structured data coming from electronic medical record databases and unstructured text coming from physician notes and published literature—to provide evidence for faster and better decision making. First, the physician and the patient can describe symptoms and other related factors to the system in natural language. Watson can then identify the key pieces of information and mine the patient's data to find relevant facts about family history, current medications, and other existing conditions. It can then combine that information with current findings from tests and then can form and test hypotheses for potential diagnoses by examining a variety of data sources—treatment guidelines, electronic medical record data, doctors' and nurses' notes, and peer-reviewed research and clinical studies. Next, Watson can suggest potential diagnostics and treatment options with a confidence rating for each suggestion.

Watson also has the potential to transform healthcare by intelligently synthesizing fragmented research findings published in a variety of outlets. It can dramatically change the way medical students learn. It can help healthcare managers to be proactive about upcoming demand patterns, optimally allocate resources, and improve processing of payments. Early examples of leading healthcare providers that use Watson-like cognitive systems include MD Anderson, The Cleveland Clinic, and Memorial Sloan Kettering.

SECURITY As the Internet expands into every facet of our lives—e-commerce, e-business, smart grids for energy, smart homes for remote control of residential gadgets and appliances—to make things easier to manage, it also opens up the potential for ill-intended people to intrude in our lives. We need smart systems like Watson that are capable of constantly monitoring for abnormal behavior and, when it is identified, preventing people from accessing our lives and harming us. This could be at the corporate or even national security system level; it could also be at the personal level. Such a smart system could learn who we are and become a digital guardian that could make inferences about activities related to our life and alert us whenever abnormal things happen.

FINANCE The financial services industry faces complex challenges. Regulatory measures as well as social and governmental pressures for financial institutions to be more inclusive have increased. And the customers the industry serves are more empowered, demanding, and sophisticated than ever before. With so much financial information generated each day, it is difficult to properly harness the appropriate information on which to act. Perhaps the solution is to create smarter client engagement by better understanding risk profiles and the operating environment. Major financial institutions are already working with Watson to infuse intelligence into their business processes. Watson is tackling data-intensive challenges across the financial services sector, including banking, financial planning, and investing.

RETAIL The retail industry is rapidly changing according to customers' needs and wants. Empowered by mobile devices and social networks that give them easier access to more information faster than ever before, customers have high expectations for products and services. While retailers are using analytics to keep up with those expectations, their bigger challenge is efficiently and effectively analyzing the growing mountain of real-time insights that could give them a competitive advantage. Watson's cognitive computing capabilities related to analyzing massive amounts of unstructured data can help retailers reinvent their decision-making processes around pricing, purchasing, distribution, and staffing. Because of Watson's ability to understand and answer questions in natural language, Watson is an effective and scalable solution for analyzing and responding to social sentiment based on data obtained from social interactions, blogs, and customer reviews.

EDUCATION With the rapidly changing characteristics of students—who are more visually oriented/stimulated, constantly connected to social media and social networks, and with increasingly shorter attention spans—what should the future of education and the classroom look like? The next generation of educational systems should be tailored to fit the needs of the new generation with customized learning plans, personalized textbooks (digital ones with integrated multimedia—audio, video, animated graphs/charts, etc.), dynamically adjusted curriculum, and perhaps smart digital tutors and 24/7 personal advisors. Watson seems to have what it takes to make all this happen. With its NLP capability, students can converse with it just as they do with their teachers, advisors, and friends.

This smart assistant can answer students' questions, satisfy their curiosity, and help them keep up with the endeavors of the educational journey.

GOVERNMENT For local, regional, and national governments, the exponential rise of Big Data presents an enormous dilemma. Today's citizens are more informed and empowered than ever before, and that means they have high expectations for the value of the public sector serving them. And government organizations can now gather enormous volumes of unstructured, unverified data that could serve their citizens, but only if those data can be analyzed efficiently and effectively. IBM Watson's cognitive computing may help make sense of this data deluge, speeding governments' decision-making processes and helping public employees to focus on innovation and discovery.

RESEARCH Every year, hundreds of billions of dollars are spent on research and development, most of it documented in patents and publications, creating an enormous amount of unstructured data. To contribute to the extant body of knowledge, one needs to sift through these data sources to find the outer boundaries of research in a particular field. This is very difficult, if not impossible, work if it is done with traditional means, but Watson can act as a research assistant to help collect and synthesize information to keep people updated on recent findings and insights. For instance, the New York Genome Center is using the IBM Watson cognitive computing system to analyze the genomic data of patients diagnosed with a highly aggressive and malignant brain cancer and to more rapidly deliver personalized, life-saving treatment to patients with this disease (Royyuru, 2014).

SECTION 6.10 REVIEW QUESTIONS

1. What is cognitive computing, and how does it differ from other computing paradigms?
2. Draw a diagram and explain the conceptual framework of cognitive computing. Make sure to include inputs, enablers, and expected outcomes in your framework.
3. List and briefly define the key attributes of cognitive computing.
4. How does cognitive computing differ from ordinary AI techniques?
5. What are the typical use cases for cognitive analytics?
6. Explain what the terms *cognitive analytics* and *cognitive search* mean.
7. What is IBM Watson and what is its significance to the world of computing?
8. How does Watson work?
9. List and briefly explain five use cases for IBM Watson.

Chapter Highlights

- Deep learning is among the latest trends in AI that come with great expectations.
- The goal of deep learning is similar to those of the other machine-leaning methods, which is to use sophisticated mathematical algorithms to learn from data similar to the way that humans learn.
- What deep learning has added to the classic machine-learning methods is the ability to automatically acquire the features required to accomplish highly complex and unstructured tasks.
- Deep learning belongs to the representation learning within the AI learning family of methods.
- The recent emergence and popularity of deep learning can largely be attributed to very large data sets and rapidly advancing commuting infrastructures.
- Artificial neural networks emulate the way the human brain works. The basic processing unit is a neuron. Multiple neurons are grouped into layers and linked together.

- In a neural network, knowledge is stored in the weight associated with the connections between neurons.
- Backpropagation is the most popular learning paradigm of feedforward neural networks.
- An MLP-type neural network consists of an input layer, an output layer, and a number of hidden layers. The nodes in one layer are connected to the nodes in the next layer.
- Each node at the input layer typically represents a single attribute that may affect the prediction.
- The usual process of learning in a neural network involves three steps: (1) compute temporary outputs based on inputs and random weights, (2) compute outputs with desired targets, and (3) adjust the weights and repeat the process.
- Developing neural network–based systems requires a step-by-step process. It includes data preparation and preprocessing, training and testing, and conversion of the trained model into a production system.
- Neural network software allows for easy experimentation with many models. Although neural network modules are included in all major data mining software tools, specific neural network packages are also available.
- Neural network applications abound in almost all business disciplines as well as in virtually all other functional areas.
- Overfitting occurs when neural networks are trained for a large number of iterations with relatively small data sets. To prevent overfitting, the training process is controlled by an assessment process using a separate validation data set.
- Neural networks are known as *black-box models*. Sensitivity analysis is often used to shed light into the black box to assess the relative importance of input features.
- Deep neural networks broke the generally accepted notion of "no more than two hidden layers are needed to formulate complex prediction problems." They promote increasing the hidden layer to arbitrarily large numbers to better represent the complexity in the data set.
- MLP deep networks, also known as *deep feedforward networks*, are the most general type of deep networks.
- The impact of random weights in the learning process of deep MLP is shown to be a significant issue. Nonrandom assignment of the initial weights seems to significantly improve the learning process in deep MLP.
- Although there is no generally accepted theoretical basis for this, it is believed and empirically shown that in deep MLP networks, multiple layers perform better and converge faster than few layers with many neurons.
- CNNs are arguably the most popular and most successful deep learning methods.
- CNNs were initially designed for computer vision applications (e.g., image processing, video processing, text recognition) but also have been shown to be applicable to nonimage or non-text data sets.
- The main characteristic of the convolutional networks is having at least one layer involving a convolution weight function instead of general matrix multiplication.
- The convolution function is a method to address the issue of having too many network weight parameters by introducing the notion of parameter sharing.
- In CNN, a convolution layer is often followed by another layer known as the *pooling* (a.k.a. *subsampling*) layer. The purpose of a pooling layer is to consolidate elements in the input matrix in order to produce a smaller output matrix while maintaining the important features.
- ImageNet is an ongoing research project that provides researchers with a large database of images, each linked to a set of synonym words (known as *synset*) from WordNet (a word hierarchy database).
- AlexNet is one of the first convolutional networks designed for image classification using the ImageNet data set. Its success rapidly popularized the use and reputation of CNNs.
- GoogLeNet (a.k.a. *Inception*), a deep convolutional network architecture designed by Google researchers, was the winning architecture at ILSVRC 2014.
- Google Lens is an app that uses deep learning artificial neural network algorithms to deliver information about the images captured by users from their nearby objects.
- Google's word2vec project remarkably increased the use of CNN-type deep learning for text mining applications.
- RNN is another deep learning architecture designed to process sequential inputs.
- RNNs have memory to remember previous information in determining context-specific, time-dependent outcomes.
- A variation of RNN, the LSTM network is today known as the most effective sequence modeling

technique and is the base of many practical applications.
- Two emerging LSTM applications are Google Neural Machine Translator and Microsoft Skype Translator.
- Deep learning implementation frameworks include Torch, Caffe, TensorFlow, Theano, and Keras.
- Cognitive computing makes a new class of problems computable by addressing highly complex situations that are characterized by ambiguity and uncertainty; in other words, it handles the kinds of problems that are thought to be solvable by human ingenuity and creativity.
- Cognitive computing finds and synthesizes data from various information sources and weighs the context and conflicting evidence inherent in the data in order to provide the best possible answers to a given question or problem.
- The key attributes of cognitive computing include adaptability, interactivity, being iterative, stateful, and contextual.
- *Cognitive analytics* is a term that refers to cognitive computing–branded technology platforms, such as IBM Watson, that specialize in the processing and analysis of large unstructured data sets.
- Cognitive search is the new generation of search method that uses AI (advanced indexing, NLP, and machine learning) to return results that are much more relevant to the user than traditional search methods.
- IBM Watson is perhaps the smartest computer system built to date. It has coined and popularized the term *cognitive computing*.
- IBM Watson beat the best of men (the two most winning competitors) at the quiz game *Jeopardy!*, showcasing the ability of commuters to do tasks that are designed for human intelligence.
- Watson and systems like it are now in use in many application areas including healthcare, finance, security, and retail.

Key Terms

activation function
artificial intelligence (AI)
artificial neural networks (ANN)
backpropagation
black-box syndrome
Caffe
cognitive analytics
cognitive computing
cognitive search
connection weight
constant error carousel (CEC)
convolution function
convolutional neural network (CNN)
deep belief network (DBN)
deep learning
deep neural network
DeepQA
Google Lens
GoogLeNet
Google Neural Machine Translator (GNMT)
graphics processing unit (GPU)
hidden layer
IBM Watson
ImageNet
Keras
long short-term memory (LSTM)
machine learning
Microsoft Skype Translator
multilayer perceptron (MLP)
MYCIN
network structure
neural network
neuron
overfitting
perceptron
performance function
pooling
processing element (PE)
recurrent neural network (RNN)
representation learning
sensitivity analysis
stochastic gradient descent (SGD)
summation function
supervised learning
TensorFlow
Theano
threshold value
Torch
transfer function
word embeddings
word2vec

Questions for Discussion

1. What is deep learning? What can deep learning do that traditional machine-learning methods cannot?
2. List and briefly explain different learning paradigms/methods in AI.
3. What is representation learning, and how does it relate to machine learning and deep learning?
4. List and briefly describe the most commonly used ANN activation functions.
5. What is MLP, and how does it work? Explain the function of summation and activation weights in MLP-type ANN.
6. List and briefly describe the nine-step process in conducting a neural network project.

7. Draw and briefly explain the three-step process of learning in ANN.
8. How does the backpropagation learning algorithm work?
9. What is overfitting in ANN learning? How does it happen, and how can it be prevented?
10. What is the so-called black-box syndrome? Why is it important to be able to explain an ANN's model structure?
11. How does sensitivity analysis work in ANN? Search the Internet to find other methods to explain ANN methods.
12. What is meant by "deep" in deep neural networks? Compare deep neural network to shallow neural network.
13. What is GPU? How does it relate to deep neural networks?
14. How does a feedforward multilayer perceptron–type deep network work?
15. Comment on the impact of random weights in developing deep MLP.
16. Which strategy is better: more hidden layers versus more neurons?
17. What is CNN?
18. For what type of applications can CNN be used?
19. What is the convolution function in CNN, and how does it work?
20. What is pooling in CNN? How does it work?
21. What is ImageNet, and how does it relate to deep learning?
22. What is the significance of AlexNet? Draw and describe its architecture.
23. What is GoogLeNet? How does it work?
24. How does CNN process text? What is word embeddings, and how does it work?
25. What is word2vec, and what does it add to the traditional text mining?
26. What is RNN? How does it differ from CNN?
27. What is the significance of *context, sequence*, and *memory* in RNN?
28. Draw and explain the functioning of a typical recurrent neural network unit.
29. What is LSTM network, and how does it differ from RNNs?
30. List and briefly describe three different types of LSTM applications.
31. How do Google's Neural Machine Translation and Microsoft Skype Translator work?
32. Despite its short tenure, why do you think deep learning implementation has several different computing frameworks?
33. Define and comment on the relationship between CPU, NVIDIA, CUDA, and deep learning.
34. List and briefly define the characteristics of different deep learning frameworks.
35. What is Keras, and how does it differ from other frameworks?
36. What is cognitive computing and how does it differ from other computing paradigms?
37. Draw a diagram and explain the conceptual framework of cognitive computing. Make sure to include inputs, enablers, and expected outcomes in your framework.
38. List and briefly define the key attributes of cognitive computing.
39. How does cognitive computing differ from ordinary AI techniques?
40. What are the typical use cases for cognitive analytics?
41. What is cognitive analytics? What is cognitive search?
42. What is IBM Watson, and what is its significance to the world of computing?
43. How does IBM Watson work?
44. List and briefly explain five use cases for IBM Watson.

Exercises

Teradata University Network (TUN) and Other Hands-On and Internet Exercises

1. Go to the Teradata University Network Web site (**teradatauniversitynetwork.com**). Search for teaching and learning materials (e.g., articles, application cases, white papers, videos, exercises) on deep learning, cognitive computing, and IBM Watson. Read the material you have found. If needed, also conduct a search on the Web to enhance your findings. Write a report on your findings.
2. Deep learning is relatively new to the world of analytics. Its application cases and success stories are just starting to emerge in the Web. Conduct a comprehensive search on your school's digital library resources to identify at least five journal articles where interesting deep learning applications are described. Write a report on your findings.
3. Most of the applications of deep learning today are developed using R- and/or Python-based open-source computing resources. Identify those resources (frameworks such as Torch, Caffe, TensorFlow, Theano, Keras) available for building deep learning models and applications. Compare and contrast their capabilities and limitations. Based on your findings and understanding of these resources, if you were to develop a deep learning application, which one would you choose to employ? Explain and justify/defend your choice.
4. *Cognitive computing* has become a popular term to define and characterize the extent of the ability of machines/computers to show "intelligent" behavior. Thanks to IBM

Watson and its success on *Jeopardy!*, cognitive computing and cognitive analytics are now part of many real-world intelligent systems. In this exercise, identify at least three application cases where cognitive computing was used to solve complex real-world problems. Summarize your findings in a professionally organized report.

5. Download KNIME analytics platform, one of the most popular free/open-source software tools from knime.org. Identify the deep learning examples (where Keras is used to build some exemplary prediction/classification models) in its example folder. Study the models in detail. Understand what it does and how exactly it does it. Then, using a different but similar data set, build and test your own deep learning prediction model. Report your findings and experiences in a written document.
6. Search for articles related to "cognitive search." Identify at least five pieces of written material (a combination of journal articles, white papers, blog posts, application cases, etc.). Read and summarize your findings. Explain your understanding of cognitive search and how it differs from regular search methods.
7. Go to **Teradata.com**. Search and find application case studies and white papers on deep learning and/or cognitive computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.
8. Go to **SAS.com**. Search and find application case studies and white papers on deep learning and/or cognitive computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.
9. Go to **IBM.com**. Search and find application case studies and white papers on deep learning and/or cognitive computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.
10. Go to **TIBCO.com** or some other advanced analytics company Web site. Search and find application case studies and white papers on deep learning and/or cognitive computing. Write a report to summarize your findings, and comment on the capabilities and limitations (based on your understanding) of these technologies.

References

Abad, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, . . . M. Isard. (2016). "TensorFlow: A System for Large-Scale Machine Learning." *OSDI, 16*, pp. 265–283.

Altman, E. I. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *The Journal of Finance, 23*(4), pp. 589–609.

Bahdanau, D., K. Cho, & Y. Bengio. (2014). "Neural Machine Translation by Jointly Learning to Align and Translate." ArXiv Preprint ArXiv:1409.0473.

Bengio, Y. (2009). "Learning Deep Architectures for AI." *Foundations and Trends® in Machine Learning, 2*(1), pp. 1–127.

Bergstra, J., O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, . . . Y. Bengio. (2010). "Theano: A CPU and GPU Math Compiler in Python." *Proceedings of the Ninth Python in Science Conference*, Vol. 1.

Bi, R. (2014). "When Watson Meets Machine Learning." **www.kdnuggets.com/2014/07/watson-meets-machine-learning.html** (accessed June 2018).

Boureau, Y.-L., N. Le Roux, F. Bach, J. Ponce, & Y. LeCun (2011). "Ask the Locals: Multi-Way Local Pooling for Image Recognition." *Proceedings of the International Computer Vision (ICCV'11) IEEE International Conference*, pp. 2651–2658.

Boureau, Y.-L., J. Ponce, & Y. LeCun. (2010). "A Theoretical Analysis of Feature Pooling in Visual Recognition." *Proceedings of International Conference on Machine Learning (ICML'10)*, pp. 111–118.

Buchanan, B. G., & E. H. Shortliffe. (1984). *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.

Cognitive Computing Consortium. (2018). **https://cognitivecomputingconsortium.com/resources/cognitive-computing-defined/#1467829079735-c0934399-599a** (accessed July 2018).

Chen, T., M. Li, Y. Li, M. Lin, N. Wang, M. Wang, . . . Z. Zhang. (2015). "Mxnet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems." ArXiv Preprint ArXiv:1512.01274.

Collobert, R., K. Kavukcuoglu, & C. Farabet. (2011). "Torch7: A Matlab-like Environment for Machine Learning." BigLearn, NIPS workshop.

Cybenko, G. (1989). "Approximation by Superpositions of a Sigmoidal Function." *Mathematics of Control, Signals and Systems, 2*(4), 303–314.

DeepQA. (2011). "DeepQA Project: FAQ, IBM Corporation." **https://researcher.watson.ibm.com/researcher/view_group.php?id=2099** (accessed May 2018).

Delen, D., R. Sharda, & M. Bessonov, M. (2006). "Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks." *Accident Analysis & Prevention*, 38(3), 434–444.

Denyer, S. (2018, January). "Beijing Bets on Facial Recognition in a Big Drive for Total Surveillance." *The Washington Post*. **https://www.washingtonpost.com/news/world/wp/2018/01/07/feature/in-china-facial-recognition-is-sharp-end-of-a-drive-for-total-surveillance/?noredirect=on&utm_term=.e73091681b31**.

Feldman, S., J. Hanover, C. Burghard, & D. Schubmehl. (2012). "Unlocking the Power of Unstructured Data." IBM White Paper. **http://.www-01.ibm.com/software/ebusiness/jstart/downloads/unlockingUnstructuredData.pdf**. (accessed May 2018).

Ferrucci, D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, Kalyanpur, A. A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, & C. Welty. (2010). "Building Watson: An Overview of the DeepQA Project." *AI Magazine, 31*(3), pp. 59–79.

Goodfellow, I., Y. Bengio, & A. Courville. (2016). "Deep Learning." Cambridge, MA: MIT Press.

Goodfellow, I. J., D. Warde-Farley, P. Lamblin, V. Dumoulin, M. Mirza, R. Pascanu, . . . Y. Bengio. (2013). "Pylearn2: A Machine Learning Research Library." ArXiv Preprint ArXiv:1308.4214.

Graves, A. (2013). "Generating Sequences with Recurrent Neural Networks." ArXiv Preprint ArXiv:1308.0850.

Graves, A., & N. Jaitly. (2014). "Towards End-to-End Speech Recognition with Recurrent Neural Networks." *Proceedings on International Conference on Machine Learning*, pp. 1764–1772.

Graves, A., N. Jaitly, & A. Mohamed. (2013). "Hybrid Speech Recognition with Deep Bidirectional LSTM." IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278.

Graves, A., A. Mohamed, & G. Hinton. (2013). "Speech Recognition with Deep Recurrent Neural Networks." IEEE Acoustics, Speech and Signal Processing (ICASSP) International Conference, pp. 6645–6649.

Graves, A., & J. Schmidhuber. (2009). "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks." *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, pp. 545–552.

Gualtieri, M. (2017). "Cognitive Search Is the AI Version of Enterprise Search, Forrester." **go.forrester.com/blogs/17-06-12-cognitive_search_is_the_ai_version_of_enterprise_search/** (accessed July 2018).

Haykin, S. S. (2009). *Neural Networks and Learning Machines*, 3rd ed. Upper Saddle River, NJ: Prentice Hall.

He, K., X. Zhang, S. Ren, & J. Sun. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification." *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034.

Hinton, G. E., S. Osindero, & Y.-W. Teh. (2006). "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation, 18*(7), 1527–1554.

Hochreiter, S., & J. Schmidhuber (1997). "Long Short-Term Memory." *Neural Computation, 9*(8), 1735–1780.

Hornik, K. (1991). "Approximation Capabilities of Multilayer Feedforward Networks." *Neural Networks, 4*(2), 251–257.

IBM. (2011). "IBM Watson." **www.ibm.com/watson/** (accessed July 2017).

Jia, Y. (2013). "Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding." **http://Goo.Gl/Fo9YO8** (accessed June 2018).

Jia, Y., E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, . . . T. Darrell, T. (2014). "Caffe: Convolutional Architecture for Fast Feature Embedding." *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678.

Keysers, D., T. Deselaers, H. A. Rowley, L.-L. Wang, & V. Carbune. (2017). "Multi-Language Online Handwriting Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), pp. 1180–1194.

Krizhevsky, A., I. Sutskever, & G. Hinton. (2012). "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems*, pp. 1097–1105S.

Kumar, S. (2017). "A Survey of Deep Learning Methods for Relation Extraction." **http://arxiv.org/abs/1705.03645**. (accessed June 2018)

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, & L. D. Jackel. (1989). "Backpropagation Applied to Handwritten ZIP Code Recognition." *Neural Computation, 1*(4), 541–551.

Liang, X., X. Shen, J. Feng, L. Lin, & S. Yan. (2016). "Semantic Object Parsing with Graph LSTM." *European Conference on Computer Vision*. New York, NY: Springer, pp. 125–143.

Mahajan, D., R. Girshick, V. Ramanathan, M. Paluri, & L. van der Maaten. (2018). "Advancing State-of-the-Art Image Recognition with Deep Learning on Hashtags." **https://code.facebook.com/posts/1700437286678763/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags/**. (accessed June 2018)

Mikolov, T., K. Chen, G. Corrado, & J. Dean. (2013). "Efficient Estimation of Word Representations in Vector Space." ArXiv Preprint ArXiv:1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, & J. Dean. (2013). "Distributed Representations of Words and Phrases and Their Compositionality" *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Mintz, M., S. Bills, R. Snow, & D. Jurafsky. (2009). "Distant Supervision for Relation Extraction Without Labeled Data." *Proceedings of the Joint Conference of the Forty-Seventh Annual Meeting of the Association for Computational Linguistics and the Fourth International Joint Conference on Natural Language Processing of the AFNLP*, Vol. 2, pp. 1003–1011.

Mozur, P. (2018, June 8). "Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras." *The New York Times*, issue June 8, 2018.

Nguyen, T. H., & R. Grishman. (2015). "Relation Extraction: Perspective from Convolutional Neural Networks." *Proceedings of the First Workshop on Vector Space Modeling for Natural Language Processing*, pp. 39–48.

Olson, D. L., D. Delen, and Y. Meng. (2012). "Comparative Analysis of Data Mining Models for Bankruptcy Prediction." *Decision Support Systems, 52*(2), pp. 464–473.

Principe, J. C., N. R. Euliano, and W. C. Lefebvre. (2000). *Neural and Adaptive Systems: Fundamentals Through Simulations*. New York: Wiley.

Reynolds, H., & S. Feldman. (2014, July/August). "Cognitive Computing: Beyond the Hype." *KM World, 23*(7), p. 21.

Riedel, S., L. Yao, & A. McCallum. (2010). "Modeling Relations and Their Mentions Without Labeled Text." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.*, New York, NY: Springer, pp. 148–163

Robinson, A., J. Levis, & G. Bennett. (2010, October). "Informs to Officially Join Analytics Movement." *ORMS Today*.

Royyuru, A. (2014). "IBM's Watson Takes on Brain Cancer: Analyzing Genomes to Accelerate and Help Clinicians Personalize Treatments." Thomas J. Watson Research Center, **www.research.ibm.com/articles/genomics.shtml** (accessed September 2014).

Rumelhart, D. E., G. E. Hinton, & R. J. Williams. (1986). "Learning Representations by Back-Propagating Errors." *Nature, 323*(6088), pp. 533.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, . . . M. Bernstein. (2015). "Imagenet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision, 115*(3), 211–252.

Sato, K., C. Young, & D. Patterson. (2017). "An In-Depth Look at Google's First Tensor Processing Unit (TPU)." **https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu**. (accessed June 2018)

Scherer, D., A. Müller, & S. Behnke. (2010). "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition." *International Conference on Artificial Neural Networks.*, New York, NY: Springer, 92–101.

Slowey, L. (2017a, January 25). "Winning the Best Picture Oscar: IBM Watson and Winning Predictions." **https://www.ibm.com/blogs/internet-of-things/best-picture-oscar-watson-predicts/**(accessed August 2018).

Slowey, L. (2017b, May 10). "Watson Predicts the Winners: Eurovision 2017." **https://www.ibm.com/blogs/internet-of-things/eurovision-watson-tone-predictions/**(accessed August 2018).

Sutskever, I., O. Vinyals, & Q. V. Le. (2014). "Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, pp. 3104–3112.

Ung, G. M. (2016, May). "Google's Tensor Processing Unit Could Advance Moore's Law 7 Years into the Future." *PC-World*. **https://www.pcworld.com/article/3072256/google-io/googles-tensor-processing-unit-said-to-advance-moores-law-seven-years-into-the-future.html (accessed** July **2018)**.

Vinyals, O., L. Kaiser, T. Koo, S. Petrov, I. Sutskever, & G. Hinton, G. (2015). "Grammar As a Foreign Language." *Advances in Neural Information Processing Systems*, pp. 2773–2781.

Vinyals, O., A. Toshev, S. Bengio, & D. Erhan. (2015). "Show and Tell: A Neural Image Caption Generator." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164.

Vinyals, O., A. Toshev, S. Bengio, & D. Erhan. (2017). "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge." *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 652–663.

Wilson, R. L., & R. Sharda. (1994). "Bankruptcy Prediction Using Neural Networks." *Decision Support Systems, 11*(5), 545–557.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, & K. Macherey. (2016). "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation." ArXiv Preprint ArXiv:1609.08144.

Xu, K., J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, & Y. Bengio. (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." *Proceedings of the Thirty-Second International Conference on Machine Learning*, pp. 2048–2057.

Zeng, D., K. Liu, S. Lai, G. Zhou, & J. Zhao (2014). "Relation Classification via Convolutional Deep Neural Network." **http://doi.org/http://aclweb.org/anthology/C/C14/C14-1220.pdf**. (accessed June 2018).

Zhou, Y.-T., R. Chellappa, A. Vaid, & B. K. Jenkins. (1988). "Image Restoration Using a Neural Network." *IEEE Transactions on Acoustics, Speech, and Signal Processing, 36*(7), pp. 1141–1151.

CHAPTER 7

Text Mining, Sentiment Analysis, and Social Analytics

LEARNING OBJECTIVES

- Describe text analytics and understand the need for text mining
- Differentiate among text analytics, text mining, and data mining
- Understand the different application areas for text mining
- Know the process of carrying out a text mining project
- Appreciate the different methods to introduce structure- to text-based data
- Describe sentiment analysis
- Develop familiarity with popular applications of sentiment analysis
- Learn the common methods for sentiment analysis
- Become familiar with speech analytics as it relates to sentiment analysis
- Learn three facets of Web analytics—content, structure, and usage mining
- Know social analytics including social media and social network analyses

This chapter provides a comprehensive overview of text analytics/mining and Web analytics/mining along with their popular application areas such as search engines, sentiment analysis, and social network/media analytics. As we have been witnessing in recent years, the unstructured data generated over the Internet of Things (IoT) (Web, sensor networks, radio-frequency identification [RFID]–enabled supply chain systems, surveillance networks, etc.) are increasing at an exponential pace, and there is no indication of its slowing down. This changing nature of data is forcing organizations to make text and Web analytics a critical part of their business intelligence/analytics infrastructure.

7.1 OPENING VIGNETTE: Amadori Group Converts Consumer Sentiments into Near-Real-Time Sales

BACKGROUND

Amadori Group, or Gruppo Amadori in Italian, is a leading manufacturing company in Italy that produces and markets food products. Headquartered in San Vittore di Cesena, Italy, the company employs more than 7,000 people and operates 16 production plants.

Amadori wanted to evolve its marketing to dynamically align with the changing lifestyles and dietary needs of young people aged 25–35. It sought to create fun ways to engage this target segment by exploiting the potential of online marketing and social media. The company wanted to boost brand visibility, encourage customer loyalty, and gauge consumers' reactions to products and marketing campaigns.

ENGAGING YOUNG ADULTS WITH CREATIVE DIGITAL MARKETING PROMOTIONS

Together with Tecla (a digital business company), Amadori used IBM WebSphere® Portal and IBM Web Content Manager software to create and manage interactive content for four mini websites, or "minisites," which promote ready-to-eat and quick-to-prepare products that fit young adults' preferences and lifestyles. For example, to market its new Evviva sausage product, the company created the "Evviva Il Würstel Italiano" minisite and let consumers upload images and videos of themselves attending events organized by Amadori. To encourage participation, the company offered the winner a spot in its next national ad campaign.

With this and other campaigns, the Amadori marketing staff compiled a database of consumer profiles by asking minisite visitors to share data to enter competitions, download applications, receive regular newsletters, and sign up for events. Additionally, the company uses Facebook Insights technology to obtain metrics on its Facebook page, including the number of new fans and favorite content.

MONITORING MARKETPLACE PERCEPTIONS OF THE AMADORI BRAND

The company capitalizes on IBM SPSS® Data Collection software to help assess peoples' opinions of its products and draw conclusions about any fluctuation in Amadori brands' popularity among consumers. For example, as it launched its Evviva campaign

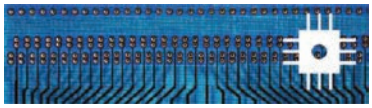
	**Instrumented**	The interactive digital platform supports rapid, accurate data collection from business partners and customers.
	**Interconnected**	The digital platform also provides an integrated view of the company's end-to-end processes from production plan to marketing and sales.
	**Intelligent**	Content management, data collection, and predictive analytics applications monitor and analyze social media relevant to the Amadori brand, helping the company anticipate issues and better align products and marketing promotions with customers' needs and desires.

FIGURE 7.1 **Smarter Commerce—Improving Consumer Engagement through Analytics.**

TV advertisements and Beach Party Tour, Amadori experienced a flood of consumer conversation. The company captured comments about the product from its Web site and social media networks using SPSS software's sentiment analysis functionality and successfully adjusted its marketing efforts in near-real time. The software does not depend solely on keyword searches but also analyzes the syntax of languages, connotations, and even slang to reveal hidden speech patterns that help gauge whether comments about the company or Amadori products express a positive, negative, or neutral opinion. Figure 7.1 shows Amadori's three facets of commerce analytics to improve consumer engagement.

MAINTAINING BRAND INTEGRITY AND CONSISTENCY ACROSS PRODUCT LINES

Building on the success of marketing minisites, Amadori launched a new corporate Web site built on the same IBM portal and content management technology. The company now concentrates on bringing visitors to the corporate Web site. Instead of individual minisites, Amadori offers sections within the Web site, some with different templates and graphics and a user interface specific to a particular marketing or ad campaign. "For example, we introduced a new product that is made from organic, free-range chickens," says Fabbri. "As part of the marketing plan, we offered webcam viewing in a new section of our corporate site so visitors could see how the poultry live and grow. We created a new graphic, but the URL, the header and the footer are always the same so visitors understand that they are always in the Amadori site."

Visitors can move from one section to another, remaining longer and learning about other offerings. With new content added weekly, the Amadori site has become bigger and is gaining greater prominence on Google and other search engines. "The first year after implementation, our Website traffic grew to approximately 240,000 unique visitors, with 30 percent becoming loyal users," says Fabbri.

KEEPING CONTENT CURRENT AND ENGAGING TO DIVERSE AUDIENCES

With more content and a high volume of traffic on the Web, it is important that visitors continue to easily find what they are looking for no matter how they access the Amadori site. With this aim in mind, the Amadori project team created a content taxonomy organized

by role and area of interest. For example, when people visit the Amadori Web site, they see a banner inviting them to "Reorganize the contents." They can identify themselves as consumers, buyers, or journalists/bloggers and slide selection bars to indicate interest level in corporate, cooking, and/or entertainment information. The content appearing on the site changes in real time based on these selections. "If the visitor identifies himself as a professional buyer interested predominantly in corporate information, the icons he sees at the top of the screen invite him to either view a digital product catalog online or download a PDF," says Fabbri. "In that same area of the screen, a consumer interested in cooking sees an icon that clicks through to pages with recipes for preparing dishes using Amadori products."

Amadori's advanced analytics projects have been producing significant business benefits, making a very strong case for the company to venture into more innovative use of social data. Following are a few of the most prevalent ones:

- Boost by 100 percent the company's ability to dynamically monitor and learn about the health of its brand using sentiment analysis.
- Improve the company's social media presence by 100 percent using near-real-time marketing insights, gaining 45,000 Facebook fans in less than one year.
- Establish direct communication with the target segment through Web integration with social media.
- Increase sales by facilitating timely promotions such as eCoupons.

As this case illustrates, in this age of Internet and social media, customer-focused companies are in a race to better communicate with their customers to obtain an intimate understanding of their needs, wants, likes, and dislikes. Social analytics that builds on social media—providing both content and the social network–related data—enables these companies to gain deeper insights than ever before.

QUESTIONS FOR THE OPENING VIGNETTE

1. According to the vignette and based on your opinion, what are the challenges that the food industry is facing today?
2. How can analytics help businesses in the food industry to survive and thrive in this competitive marketplace?
3. What were and still are the main objectives for Amadori to embark into analytics? What were the results?
4. Can you think of other businesses in the food industry that utilize analytics to become more competitive and customer focused? If not, an Internet search could help find relevant information to answer this question.

WHAT WE CAN LEARN FROM THIS VIGNETTE

It is safe to say that computer technology, both on the hardware and software fronts, is advancing faster than anything else in the last 50-plus years. Things that were too big, too complex, and impossible to solve are now well within the reach of information technology. One of the enabling technologies is perhaps text analytics/text mining and its derivative called *sentiment analysis*. Traditionally, we have created databases to structure the data so that they can be processed by computers. Textual content, on the other hand, has always been meant for humans to process. Can machines do the things that are meant for humans' creativity and intelligence? Evidently, yes! This case illustrates the viability and value proposition of collecting and processing customer opinions to develop new and improved products and services, managing the company's brand name, and engaging and energizing the customer base for mutually beneficial and closer relationships. Under the overarching name of "digital marketing," Amadori showcases the use of text

mining, sentiment analysis, and social media analytics to significantly advance the bottom line through improved customer satisfaction, increased sales, and enhanced brand loyalty.

Sources: IBM Customer Case Study. "Amadori Group Converts Consumer Sentiments into Near-Real-Time Sales." Used with permission of IBM.

7.2 TEXT ANALYTICS AND TEXT MINING OVERVIEW

The information age that we are living in is characterized by the rapid growth in the amount of data and information collected, stored, and made available in electronic format. A vast majority of business data are stored in text documents that are virtually unstructured. According to a study by Merrill Lynch and Gartner, 85 percent of all corporate data are captured and stored in some sort of unstructured form (McKnight, 2005). The same study also stated that these unstructured data are doubling in size every 18 months. Because knowledge is power in today's business world and knowledge is derived from data and information, businesses that effectively and efficiently tap into their text data sources will have the necessary knowledge to make better decisions, leading to a competitive advantage over those businesses that lag behind. This is where the need for text analytics and text mining fits into the big picture of today's businesses.

Even though the overarching goal for both *text analytics* and *text mining* is to turn unstructured textual data into actionable information through the application of natural language processing (NLP) and analytics, the definitions of these terms are somewhat different, at least to some experts in the field. According to them, "text analytics" is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms), as well as information extraction, data mining, and Web mining, whereas "text mining" is primarily focused on discovering new and useful knowledge from the textual data sources. Figure 7.2 illustrates the relationships between text analytics and text mining along with other related application areas. The bottom of Figure 7.2 lists the main disciplines (the foundation of the house) that play a critical role in the development of these increasingly more popular application areas. Based on this definition of text analytics and text mining, one could simply formulate the difference between the two as follows:

$$\textit{Text Analytics} = \textit{Information Retrieval} + \textit{Information Extraction} + \textit{Data Mining} + \textit{Web Mining}$$

or simply

$$\textit{Text Analytics} = \textit{Information Retrieval} + \textit{Text Mining}$$

Compared to text mining, *text analytics* is a relatively new term. With the recent emphasis on *analytics*, as has been the case in many other related technical application areas (e.g., consumer analytics, completive analytics, visual analytics, social analytics), the field of text has also wanted to get on the analytics bandwagon. Although the term *text analytics* is more commonly used in a business application context, *text mining* is frequently used in academic research circles. Even though the two can be defined somewhat differently at times, *text analytics* and *text mining* are usually used synonymously, and we (authors of this book) concur with this.

Text mining (also known as *text data mining* or *knowledge discovery in textual databases*) is the semiautomated process of extracting patterns (useful information and knowledge) from large amounts of unstructured data sources. Remember that data mining is the process of identifying valid, novel, potentially useful, and ultimately understandable

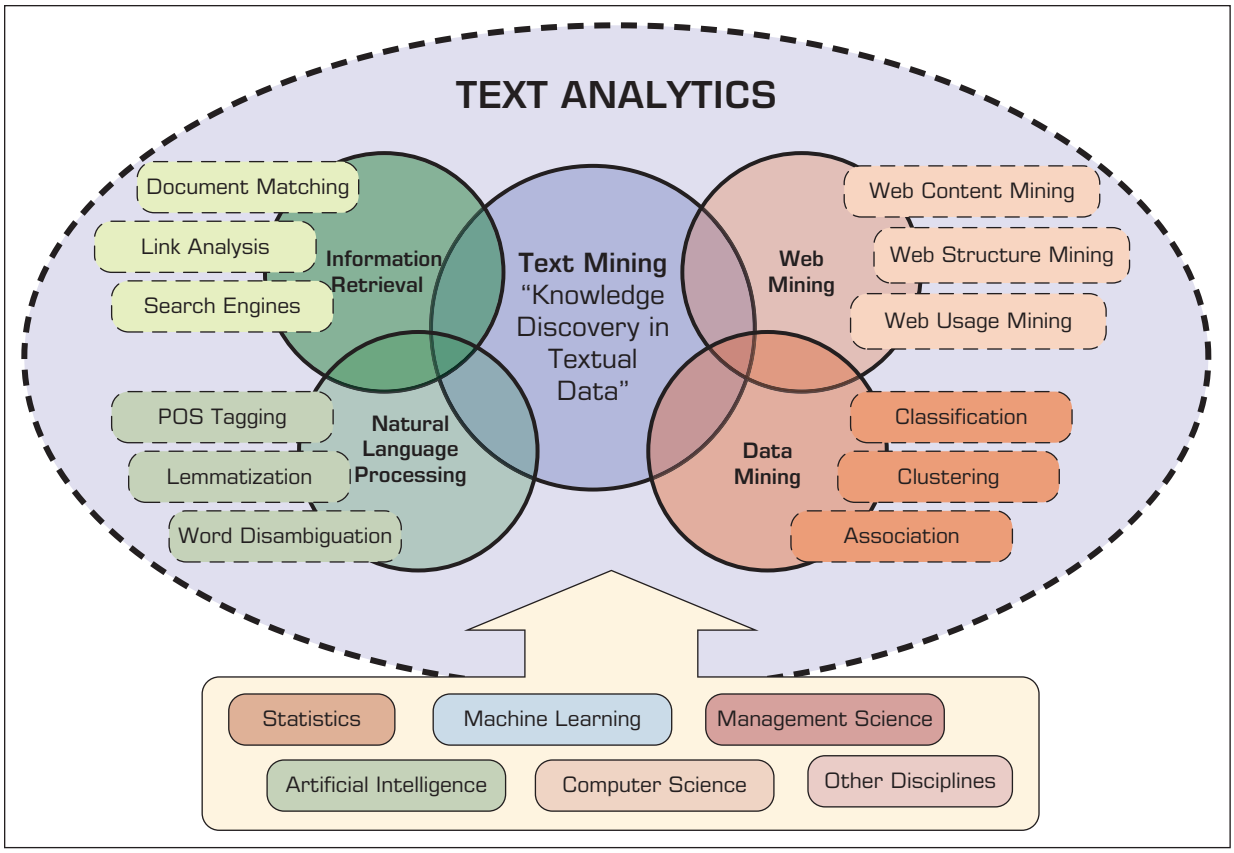


FIGURE 7.2 **Text Analytics, Related Application Areas, and Enabling Disciplines.**

patterns in data stored in structured databases where the data are organized in records structured by categorical, ordinal, or continuous variables. Text mining is the same as data mining in that it has the same purpose and uses the same processes, but with text mining, the input to the process is a collection of unstructured (or less structured) data files such as Word documents, PDF files, text excerpts, and XML files. In essence, text mining can be thought of as a process (with two main steps) that starts with imposing structure on the text-based data sources followed by extracting relevant information and knowledge from these structured text-based data using data mining techniques and tools.

The benefits of text mining are obvious in the areas in which very large amounts of textual data are being generated, such as law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), and marketing (customer comments). For example, the free-form text-based interactions with customers in the form of complaints (or compliments) and warranty claims can be used to objectively identify product and service characteristics that are deemed to be less than perfect and can be used as input to better product development and service allocations. Likewise, market outreach programs and focus groups generate large amounts of data. By not restricting product or service feedback to a codified form, customers can present, in their own words, what they think about a company's products and services. Another area where the automated processing of unstructured text has had much impact is in electronic communications and e-mail. Text mining can be used not only to classify and filter junk e-mail but also to automatically

prioritize e-mail based on importance level as well as generate automatic responses (Weng & Liu, 2004). Following are among the most popular application areas of text mining:

- **Information extraction.** Identifying key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching.
- **Topic tracking.** Based on a user profile and documents that a user views, predicting other documents of interest to the user.
- **Summarization.** Summarizing a document to save the reader time.
- **Categorization.** Identifying the main themes of a document and then placing the document into a predefined set of categories based on those themes.
- **Clustering.** Grouping similar documents without having a predefined set of categories.
- **Concept linking.** Connecting related documents by identifying their shared concepts and, by doing so, helping users find information that they perhaps would not have found using traditional search methods.
- **Question answering.** Finding the best answer to a given question through knowledge-driven pattern matching.

See Technology Insights 7.1 for explanations of some of the terms and concepts used in text mining. Application Case 7.1 describes the use of text mining in the insurance industry.

Application Case 7.1 shows how text mining and a variety of user-generated data sources enable Netflix stay innovative in its business practices, generate deeper customer insight, and drive very successful content for its viewers.

TECHNOLOGY INSIGHTS 7.1 Text Mining Terminology

The following list describes some commonly used text mining terms:

- ***Unstructured data (versus structured data).*** Structured data have a predetermined format. They are usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, **unstructured data** do not have a predetermined format and are stored in the form of textual documents. In essence, structured data are for the computers to process, whereas unstructured data are for humans to process and understand.
- ***Corpus.*** In linguistics, a **corpus** (plural *corpora*) is a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.
- ***Terms.*** A *term* is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of NLP methods.
- ***Concepts.*** *Concepts* are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are the result of higher-level abstraction.
- ***Stemming.*** **Stemming** is the process of reducing inflected words to their stem (or base or root) form. For instance, *stemmer, stemming, stemmed* are all based on the root *stem*.
- ***Stop words.*** **Stop words** (or *noise words*) are words that are filtered out prior to or after processing natural language data (i.e., text). Even though there is no universally accepted list of stop words, most NLP tools use a list that includes articles (*a, an, the*), prepositions (*of, on, for*), auxiliary verbs (*is, are, was, were*), and context-specific words that are deemed not to have differentiating value.
- ***Synonyms and polysemes.*** Synonyms are syntactically different words (i.e., spelled differently) with identical or at least similar meanings (e.g., *movie, film*, and *motion picture*). In contrast, **polysemes**, which are also called *homonyms*, are syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., *bow* can mean "to bend forward," "the front of the ship," "the weapon that shoots arrows," or "a kind of tied ribbon").

- ***Tokenizing.*** A *token* is a categorized block of text in a sentence. The block of text corresponding to the token is categorized according to the function it performs. This assignment of meaning to blocks of text is known as **tokenizing**. A token can look like anything; it just needs to be a useful part of the structured text.
- ***Term dictionary.*** This is a collection of terms specific to a narrow field that can be used to restrict the extracted terms within a corpus.
- ***Word frequency.*** This is the number of times a word is found in a specific document.
- ***Part-of-speech tagging.*** This is the process of marking the words in a text as corresponding to a particular part of speech (nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and the context in which it is used.
- ***Morphology.*** This is the branch of the field of linguistics and a part of NLP that studies the internal structure of words (patterns of word formation within a language or across languages).
- ***Term-by-document matrix (occurrence matrix).*** This term refers to the common representation schema of the frequency-based relationship between the terms and documents in tabular format where terms are listed in columns, documents are listed in rows, and the frequency between the terms and documents is listed in cells as integer values.
- ***Singular value decomposition (latent semantic indexing).*** This dimensionality reduction method is used to transform the term-by-document matrix to a manageable size by generating an intermediate representation of the frequencies using a matrix manipulation method similar to principal component analysis.

Application Case 7.1

Netflix: Using Big Data to Drive Big Engagement: Unlocking the Power of Analytics to Drive Content and Consumer Insight

The Problem

In today's hyper-connected world, businesses are under enormous pressure to build relationships with fully engaged consumers who keep coming back for more.

In theory, fostering more intimate consumer relationships becomes easier as new sources of data emerge, data volumes continue their unprecedented growth, and technology becomes more sophisticated. These developments should enable businesses to do a much better job of personalizing marketing campaigns and generating precise content recommendations that drive engagement, adoption, and value for subscribers.

Yet achieving an advanced understanding of one's audience is a continuous process of testing and learning. It demands the ability to quickly gather and reliably analyze thousands, millions, even billions of events every day found in a variety of data sources, formats, and locations—otherwise known as Big Data. Technology platforms crafted to gather these data and conduct the analyses must be powerful enough to deliver timely insights today and flexible enough to change and grow in business and technology landscapes that morph with remarkable speed.

Netflix, an undisputed leader and innovator in the over-the-top (OTT) content space, understands this context better than most. It has staked its business and its brand on delivering highly targeted, personalized experiences for every subscriber—and has even begun using its remarkably detailed insights to change the way it buys, licenses, and develops content, causing many throughout the Media and Entertainment industries to sit up and take notice.

To support these efforts, Netflix leverages Teradata as a critical component of its data and analytics platform. More recently, the two companies partnered to transition Netflix to the Teradata Cloud, which has given Netflix the power and flexibility it needs—and, so, the ability to maintain its focus on those initiatives at the core of its business.

A Model for Data-Driven, Consumer-Focused Business

The Netflix story is a model for data-driven, direct-to-consumer, and subscriber-based companies—and, in fact, for any business that needs engaged audiences to thrive in a rapidly changing world.

(*Continued*)

Application Case 7.1 (Continued)

After beginning as a mail-order DVD business, Netflix became the first prominent OTT content provider and turned the media world on its head; witness recent decisions by other major media companies to begin delivering OTT content.

One major element in Netflix's success is the way it relentlessly tweaks its recommendation engines, constantly adapting to meet each consumer's preferred style. Most of the company's streaming activity emerges from its recommendations, which generate enormous consumer engagement and loyalty. Every interaction a Netflix subscriber has with the service is based on meticulously culled and analyzed interactions—no two experiences are the same.

In addition, as noted above, Netflix has applied its understanding of subscribers and potential subscribers—as individuals and as groups—to make strategic purchasing, licensing, and content development decisions. It has created two highly successful dramatic series—House of Cards and Orange is the New Black—that are informed in part by the company's extraordinary understanding of its subscribers.

While those efforts and the business minds that drive them make up the heart of the company's business, the technology that supports these initiatives must be more powerful and reliable than that of its competitors. The data and analytics platform must be able to:

- Rapidly and reliably handle staggering workloads; it must support insightful analysis of billions of transactional events each day—every search, browse, stop, and start—in whatever data format that records the events.
- Work with a variety of analytics approaches, including neural networks, Python, Pig, as well as varied Business Intelligence tools, like MicroStrategy.
- Easily scale and contract as necessary with exceptional elasticity.
- Provide a safe and redundant repository for all of the company's data.
- Fit within the company's cost structure and desired profit margins.

Bringing Teradata Analytics to the Cloud

With these considerations in mind, Netflix and Teradata teamed up to launch a successful venture to bring Netflix's Teradata Data Warehouse into the cloud.

Power and Maturity: Teradata's well-earned reputation for exceptional performance is especially important to a company like Netflix, which pounds its analytics platform with hundreds of concurrent queries. Netflix also needed data warehousing and analytics tools that enable complex workload management—essential for creating different queues for different users, and thus allowing for the constant and reliable filtering of what each user needs.

Hybrid Analytical Ecosystems and a Unified Data Architecture: Netflix's reliance on a hybrid analytical ecosystem that leverages Hadoop where appropriate but refuses to compromise on speed and agility was the perfect fit for Teradata. Netflix's cloud environment relies on a Teradata-Hadoop connector that enables Netflix to seamlessly move cloud-based data from another provider into the Teradata Cloud. The result is that Netflix can do much of its analytics off a world-class data warehouse in the Teradata Cloud that offers peace-of-mind redundancy, the ability to expand and contract in response to changing business conditions, and a significantly reduced need for data movement. And, Netflix's no-holds-barred approach to allowing their analysts to use whatever analytical tools fit the bill demanded a unique analytics platform that could accommodate them. Having a partner that works efficiently with the full complement of analytical applications—both its own and other leading software providers—was critical.

Teradata's Unified Data Architecture (UDA) helps provide this by recognizing that most companies need a safe, cost-effective collection of services, platforms, applications, and tools for smarter data management, processing, and analytics. In turn, organizations can get the most from all their data. The Teradata UDA includes:

- An integrated data warehouse, which enables organizations to access a comprehensive and shared data environment to quickly and reliably operationalize insights throughout an organization.
- A powerful discovery platform offers companies discovery analytics that rapidly unlock insights from all available data using a variety of techniques accessible to mainstream business analysts.

- A data platform (e.g., Hadoop) provides the means to economically gather, store, and refine all a company's data and facilitate the type of discovery never before believed possible.

The Proof Is in the Eyeballs

Netflix scrupulously adheres to a few simple and powerful metrics when evaluating the success of its personalization capabilities: eyeballs. Are subscribers watching? Are they watching more? Are they watching more of what interests them?

With engagement always top of mind, it's no surprise that Netflix is among the world's leaders in personalizing content to successfully attract and retain profitable consumers. It has achieved this standing by drawing on its understanding that in a rapidly changing business and technology landscape, one key to success is constantly testing new ways of gathering and analyzing data to deliver the most effective and targeted recommendations. Working with technology partners that make such testing possible frees Netflix to focus on its core business.

Moving ahead, Netflix believes that making increased use of cloud-based technology will further empower its customer engagement initiatives. By relying on technology partners that understand how to tailor solutions and provide peace of mind about the redundancy of Netflix's data, the company expects to continue its organic growth and expand its capacity to respond nimbly to technological change and the inevitable ebbs and flows of business.

Questions for Case 7.1

1. What does Netflix do? How did they evolve into this current business model?
2. In the case of Netflix, what was it meant to be data-driven and customer-focused?
3. How did Netflix use Teradata technologies in its analytics endeavors?

Source: Teradata Case Study "Netflix: Using Big Data to Drive Big Engagement" **https://www.teradata.com/Resources/Case-Studies/Netflix-Using-Big-Data-to-Drive-Big-Engageme** (accessed July 2018).

SECTION 7.2 REVIEW QUESTIONS

1. What is text analytics? How does it differ from text mining?
2. What is text mining? How does it differ from data mining?
3. Why is the popularity of text mining as an analytics tool increasing?
4. What are some of the most popular application areas of text mining?

7.3 NATURAL LANGUAGE PROCESSING (NLP)

Some of the early text mining applications used a simplified representation called *bag-of-words* when introducing structure to a collection of text-based documents to classify them into two or more predetermined classes or to cluster them into natural groupings. In the bag-of-words model, text, such as a sentence, paragraph, or complete document, is represented as a collection of words, disregarding the grammar or the order in which the words appear. The bag-of-words model is still used in some simple document classification tools. For instance, in spam filtering, an e-mail message can be modeled as an unordered collection of words (a bag-of-words) that is compared against two different predetermined bags. One bag is filled with words found in spam messages and the other is filled with words found in legitimate e-mails. Although some of the words are likely to be found in both bags, the "spam" bag will contain spam-related words such as *stock, Viagra*, and *buy* much more frequently than the legitimate bag, which will contain more words related to the user's friends or workplace. The level of match between a specific e-mail's bag-of-words and the two bags containing the descriptors determines the membership of the e-mail as either spam or legitimate.

Naturally, we (humans) do not use words without some order or structure. We use words in sentences, which have semantic as well as syntactic structure. Thus, automated techniques (such as text mining) need to look for ways to go beyond the bag-of-words

interpretation and incorporate more and more semantic structure into their operations. The current trend in text mining is toward including many of the advanced features that can be obtained using NLP.

It has been shown that the bag-of-words method might not produce good enough information content for text mining tasks (e.g., classification, clustering, association). A good example of this can be found in evidence-based medicine. A critical component of evidence-based medicine is incorporating the best available research findings into the clinical decision-making process, which involves appraisal of the information collected from the printed media for validity and relevance. Several researchers from the University of Maryland developed evidence assessment models using a bag-of-words method (Lin and Demner-Fushman, 2005). They employed popular machine-learning methods along with more than half a million research articles collected from Medical Literature Analysis and Retrieval System Online (MEDLINE). In their models, the researchers represented each abstract as a bag-of-words, where each stemmed term represented a feature. Despite using popular classification methods with proven experimental design methodologies, their prediction results were not much better than simple guessing, which could indicate that the bag-of-words is not generating a good enough representation of the research articles in this domain; hence, more advanced techniques such as NLP were needed.

Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics. It studies the problem of "understanding" the natural human language with the task of converting depictions of human language (such as textual documents) into more formal representations (in the form of numeric and symbolic data) that are easier for computer programs to manipulate. The goal of NLP is to move beyond syntax-driven text manipulation (which is often called *word counting*) to a true understanding and processing of natural language that considers grammatical and semantic constraints as well as the context.

The definition and scope of the word *understanding* is one of the major discussion topics in NLP. Considering that the natural human language is vague and that a true understanding of meaning requires extensive knowledge of a topic (beyond what is in the words, sentences, and paragraphs), will computers ever be able to understand natural language the same way and with the same accuracy that humans do? Probably not! NLP has come a long way from the days of simple word counting, but it has an even longer way to go to really understand natural human language. The following are just a few of the challenges commonly associated with the implementation of NLP:

- **Part-of-speech tagging.** It is difficult to mark up terms in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, or adverbs) because the part of speech depends not only on the definition of the term but also on the context within which it is used.
- **Text segmentation.** Some written languages, such as Chinese, Japanese, and Thai, do not have single-word boundaries. In these instances, the text-parsing task requires the identification of word boundaries, which is often difficult. Similar challenges in speech segmentation emerge when analyzing spoken language because sounds representing successive letters and words blend into each other.
- **Word sense disambiguation.** Many words have more than one meaning. Selecting the meaning that makes the most sense can be accomplished only by taking into account the context within which the word is used.
- **Syntactic ambiguity.** The grammar for natural languages is ambiguous; that is, multiple possible sentence structures often need to be considered. Choosing the most appropriate structure usually requires a fusion of semantic and contextual information.

- **Imperfect or irregular input.** Foreign or regional accents and vocal impediments in speech and typographical or grammatical errors in texts make the processing of the language an even more difficult task.
- **Speech acts.** A sentence can often be considered an action by the speaker. The sentence structure alone might not contain enough information to define this action. For example, "Can you pass the class?" requests a simple yes/no answer, whereas "Can you pass the salt?" is a request for a physical action to be performed.

A long-standing dream of the artificial intelligence community is to have algorithms that are capable of automatically reading and obtaining knowledge from text. By applying a learning algorithm to parsed text, researchers from Stanford University's NLP lab have developed methods that can automatically identify the concepts and relationships between those concepts in the text. By applying a unique procedure to large amounts of text, the lab's algorithms automatically acquire hundreds of thousands of items of world knowledge and use them to produce significantly enhanced repositories for WordNet. **WordNet** is a laboriously hand-coded database of English words, their definitions, sets of synonyms, and various semantic relations between synonym sets. It is a major resource for NLP applications, but it has proven to be very expensive to build and maintain manually. By automatically inducing knowledge into WordNet, the potential exists to make it an even greater and more comprehensive resource for NLP at a fraction of the cost. One prominent area in which the benefits of NLP and WordNet are already being harvested is in customer relationship management (CRM). Broadly speaking, the goal of CRM is to maximize customer value by better understanding and effectively responding to customers' actual and perceived needs. An important area of CRM in which NLP is making a significant impact is sentiment analysis. **Sentiment analysis** is a technique used to detect favorable and unfavorable opinions toward specific products and services using a large number of textual data sources (customer feedback in the form of Web postings). A detailed coverage of sentiment analysis and WordNet is given in Section 7.6.

Analytics in general and text analytics and text mining in particular can be used in the broadcasting industry. Application Case 7.2 provides an example that uses a wide range of analytics capabilities to capture new viewers, predict ratings, and add business value to a broadcasting company.

Application Case 7.2

AMC Networks Is Using Analytics to Capture New Viewers, Predict Ratings, and Add Value for Advertisers in a Multichannel World

Over the past 10 years, the cable television sector in the United States has enjoyed a period of growth that has enabled unprecedented creativity in the creation of high-quality content. AMC Networks has been at the forefront of this new golden age of television, producing a string of successful, critically acclaimed shows such as *Breaking Bad, Mad Men*, and *The Walking Dead*.

Dedicated to producing quality programming and movie content for more than 30 years, AMC Networks owns and operates several of the most popular and award-winning brands in cable television, producing and delivering distinctive, compelling, and culturally relevant content that engages audiences across multiple platforms.

Getting Ahead of the Game

Despite its success, AMC Networks has no plans to rest on its laurels. As Vitaly Tsivin, SVP Business Intelligence, explains:

> We have no interest in standing still. Although a large percentage of our business is still linear

(*Continued*)

Application Case 7.2 (Continued)

cable TV, we need to appeal to a new generation of millennials who consume content in very different ways.

TV has evolved into a multichannel, multistream business, and cable networks need to get smarter about how they market to and connect with audiences across all of those streams. Relying on traditional ratings data and third-party analytics providers is going to be a losing strategy: you need to take ownership of your data, and use it to get a richer picture of who your viewers are, what they want, and how you can keep their attention in an increasingly crowded entertainment marketplace

Zoning in on the Viewer

The challenge is that there is just so much information available—hundreds of billions of rows of data from industry data-providers such as Nielsen and comScore, from channels such as AMC's TV Everywhere live Web streaming and video-on-demand service, from retail partners such as iTunes and Amazon, and from third-party online video services such as Netflix and Hulu.

"We can't rely on high-level summaries; we need to be able to analyze both structured and unstructured data, minute-by-minute and viewer-by-viewer," says Tsivin. "We need to know who's watching and why—and we need to know it quickly so that we can decide, for example, whether to run an ad or a promo in a particular slot during tomorrow night's episode of *Mad Men*."

AMC decided it needed to develop an industry-leading analytics capability in-house and focused on delivering this capability as quickly as possible. Instead of conducting a prolonged and expensive vendor and product selection process, AMC decided to leverage its existing relationship with IBM as its trusted strategic technology partner. The time and money traditionally spent on procurement were instead invested in realizing the solution, accelerating AMC's progress on its analytics roadmap by at least six months.

Web-Based Dashboard Used by AMC Networks. *Source:* Used with permission of AMC Networks.

Empowering the Research Department

In the past, AMC's research team spent a large portion of its time processing data. Today, thanks to its new analytics tools, it is able to focus most of its energy on gaining actionable insights.

"By investing in big data analytics technology from IBM, we've been able to increase the pace and detail of our research an order of magnitude," says Tsivin. "Analyses that used to take days and weeks are now possible in minutes, or even seconds." He added,

> Bringing analytics in-house will provide major ongoing cost-savings. Instead of paying hundreds of thousands of dollars to external vendors when we need some analysis done, we can do it ourselves—more quickly, more accurately, and much more cost-effectively. We're expecting to see a rapid return on investment.
>
> As more sources of potential insight become available and analytics becomes more strategic to the business, an in-house approach is really the only viable way forward for any network that truly wants to gain competitive advantage from its data.

Driving Decisions with Data

Many of the results delivered by this new analytics capability demonstrate a real transformation in the way AMC operates. For example, the company's business intelligence department has been able to create sophisticated statistical models that help the company refine its marketing strategies and make smarter decisions about how intensively it should promote each show.

With deeper insight into viewership, AMC's direct marketing campaigns are also much more successful than before. In one recent example, intelligent segmentation and look-alike modeling helped the company target new and existing viewers so effectively that AMC video-on-demand transactions were higher than would be expected otherwise.

This newfound ability to reach out to new viewers based on their individual needs and preferences is not just valuable for AMC; it also has huge potential value for the company's advertising partners. AMC is currently working on providing access to its rich data sets and analytics tools as a service for advertisers, helping them fine-tune their campaigns to appeal to ever-larger audiences across both linear and digital channels.

Tsivin concludes, "Now that we can really harness the value of big data, we can build a much more attractive proposition for both consumers and advertisers—creating even better content, marketing it more effectively, and helping it reach a wider audience by taking full advantage of our multichannel capabilities."

Questions for Case 7.2

1. What are the common challenges that broadcasting companies are facing today? How can analytics help to alleviate these challenges?
2. How did AMC leverage analytics to enhance its business performance?
3. What were the types of text analytics and text minisolutions developed by AMC networks? Can you think of other potential uses of text mining applications in the broadcasting industry?

Sources: IBM Customer Case Study. "Using Analytics to Capture New Viewers, Predict Ratings and Add Value for Advertisers in a Multichannel World." **http://www-03.ibm.com/software/businesscasestudies/us/en/corp?synkey=A023603A76220M60** (accessed July 2016); **www.ibm.com**; **www.amcnetworks.com**.

	**Instrumented**	AMC combines ratings data with viewer information from a wide range of digital channels: its own video-on-demand and live-streaming services, retailers, and online TV services.
	**Interconnected**	A powerful and comprehensive big data and analytics engine centralizes the data and makes them available to a range of descriptive and predictive analytics tools for accelerated modeling, reporting, and analysis.
	**Intelligent**	AMC can predict which shows will be successful, how it should schedule them, what promos it should create, and to whom it should market them—helping to win new audience share in an increasingly competitive market.

NLP has successfully been applied to a variety of domains for a wide range of tasks via computer programs to automatically process natural human language that previously could be done only by humans. Following are among the most popular of these tasks:

- **Question answering.** The task of automatically answering a question posed in natural language; that is, producing a human language answer when given a human language question. To find the answer to a question, the computer program can use either a prestructured database or a collection of natural language documents (a text corpus such as the World Wide Web).
- **Automatic summarization.** The creation of a shortened version of a textual document by a computer program that contains the most important points of the original document.
- **Natural language generation.** The conversion of information from computer databases into readable human language.
- **Natural language understanding.** The conversion of samples of human language into more formal representations that are easier for computer programs to manipulate.
- **Machine translation.** The automatic translation of one human language to another.
- **Foreign language reading.** A computer program that assists a nonnative language speaker in reading a foreign language with correct pronunciation and accents on different parts of the words.
- **Foreign language writing.** A computer program that assists a nonnative language user in writing in a foreign language.
- **Speech recognition.** Conversion of spoken words to machine-readable input. Given a sound clip of a person speaking, the system produces a text dictation.
- **Text to speech.** Also called *speech synthesis*, a computer program that automatically converts normal language text into human speech.
- **Text proofing.** A computer program that reads a proof copy of a text to detect and correct any errors.
- **Optical character recognition.** The automatic translation of images of handwritten, typewritten, or printed text (usually captured by a scanner) into machine-editable textual documents.

The success and popularity of text mining depends greatly on advancements in NLP in both generating and understanding human languages. NLP enables the extraction of features from unstructured text so that a wide variety of data mining techniques can be used to extract knowledge (novel and useful patterns and relationships) from it. In that sense, simply put, text mining is a combination of NLP and data mining.

SECTION 7.3 REVIEW QUESTIONS

1. What is NLP?
2. How does NLP relate to text mining?
3. What are some of the benefits and challenges of NLP?
4. What are the most common tasks addressed by NLP?

7.4 TEXT MINING APPLICATIONS

As the amount of unstructured data collected by organizations increases, so do the value proposition and popularity of text mining tools. Many organizations are now realizing the importance of extracting knowledge from their document-based data repositories through

the use of text mining tools. The following is only a small subset of the exemplary application categories of text mining.

Marketing Applications

Text mining can be used to increase cross-selling and up-selling by analyzing the unstructured data generated by call centers. Text generated notes from call center as well as transcriptions of voice conversations with customers can be analyzed by text mining algorithms to extract novel, actionable information about customers' perceptions toward a company's products and services. In addition, blogs, user reviews of products at independent Web sites, and discussion board postings are gold mines of customer sentiments. This rich collection of information, once properly analyzed, can be used to increase satisfaction and the overall lifetime value of the customer (Coussement & Van den Poel, 2008).

Text mining has become invaluable for CRM. Companies can use text mining to analyze rich sets of unstructured text data combined with the relevant structured data extracted from organizational databases to predict customer perceptions and subsequent purchasing behavior. Coussement and Van den Poel (2009) successfully applied text mining to significantly improve a model's ability to predict customer churn (i.e., customer attrition) so that those customers identified as most likely to leave a company are accurately identified for retention tactics.

Ghani et al. (2006) used text mining to develop a system capable of inferring implicit and explicit attributes of products to enhance retailers' ability to analyze product databases. Treating products as sets of attribute–value pairs rather than as atomic entities can potentially boost the effectiveness of many business applications, including demand forecasting, assortment optimization, product recommendations, assortment comparison across retailers and manufacturers, and product supplier selection. The proposed system allows a business to represent its products in terms of attributes and attribute values without much manual effort. The system learns these attributes by applying supervised and semi-supervised learning techniques to product descriptions found on retailers' Web sites.

Security Applications

One of the largest and most prominent text mining applications in the security domain is probably the highly classified ECHELON surveillance system. As rumor has it, ECHELON is assumed to be capable of identifying the content of telephone calls, faxes, e-mails, and other types of data, intercepting information sent via satellites, public-switched telephone networks, and microwave links.

In 2007, the European Union Agency for Law Enforcement Cooperation (EUROPOL) developed an integrated system capable of accessing, storing, and analyzing vast amounts of structured and unstructured data sources to track transnational organized crime. Called the Overall Analysis System for Intelligence Support (OASIS), it aims to integrate the most advanced data and text mining technologies available in today's market. The system has enabled EUROPOL to make significant progress in supporting its law enforcement objectives at the international level (EUROPOL, 2007).

The U.S. Federal Bureau of Investigation (FBI) and the Central Intelligence Agency (CIA), under the direction of the Department for Homeland Security, are jointly developing a supercomputer data and text mining system. The system is expected to create a gigantic data warehouse along with a variety of data and text mining modules to meet the knowledge-discovery needs of federal, state, and local law enforcement agencies. Prior to this project, the FBI and CIA each had its own separate database with little or no interconnection.

Another security-related application of text mining is in the area of **deception detection**. Applying text mining to a large set of real-world criminal (person-of-interest) statements, Fuller, Biros, and Delen (2008) developed prediction models to differentiate deceptive statements from truthful ones. Using a rich set of cues extracted from textual statements, the model predicted the holdout samples with 70 percent accuracy, which is believed to be a significant success considering that the cues are extracted only from textual statements (no verbal or visual cues are present). Furthermore, compared to other deception-detection techniques, such as polygraphs, this method is nonintrusive and widely applicable to not only textual data but also (potentially) transcriptions of voice recordings. A more detailed description of text-based deception detection is provided in Application Case 7.3.

Biomedical Applications

Text mining holds great potential for the medical field in general and biomedicine in particular for several reasons. First, published literature and publication outlets (especially with the advent of the open source journals) in the field are expanding at an exponential rate. Second, compared to most other fields, medical literature is more standardized and orderly, making it a more "minable" information source. Finally, the terminology used in

Application Case 7.3 Mining for Lies

Driven by advancements in Web-based information technologies and increasing globalization, computer-mediated communication continues to filter into everyday life, bringing with it new venues for deception. The volume of text-based chat, instant messaging, text messaging, and text generated by online communities of practice is increasing rapidly. Even the use of e-mail continues to increase. With the massive growth of text-based communication, the potential for people to deceive others through computer-mediated communication has also grown, and such deception can have disastrous results.

Unfortunately, in general, humans tend to perform poorly at deception-detection tasks. This phenomenon is exacerbated in text-based communications. A large part of the research on deception detection (also known as *credibility assessment*) has involved face-to-face meetings and interviews. Yet with the growth of text-based communication, text-based deception-detection techniques are essential.

Techniques for successfully detecting deception—that is, lies—have wide applicability. Law enforcement can use decision support tools and techniques to investigate crimes, conduct security screening in airports, and monitor communications of suspected terrorists. Human resources professionals might use deception-detection tools to screen applicants. These tools and techniques also have the potential to screen e-mails to uncover fraud or other wrongdoings committed by corporate officers. Although some people believe that they can readily identify those who are not being truthful, a summary of deception research showed that, on average, people are only 54 percent accurate in making veracity determinations (Bond & DePaulo, 2006). This figure may actually be worse when humans try to detect deception in text.

Using a combination of text mining and data mining techniques, Fuller et al. (2008) analyzed person-of-interest statements completed by people involved in crimes on military bases. In these statements, suspects and witnesses are required to write their recollection of the event in their own words. Military law enforcement personnel searched archival data for statements that they could conclusively identify as being truthful or deceptive. These decisions were made on the basis of corroborating evidence and case resolution. Once labeled as truthful or deceptive, the law enforcement personnel removed identifying information and gave the statements to the research team. In total, 371 usable statements were received for analysis. The text-based deception-detection method used by Fuller et al. was based on a process known as *message feature mining*, which relies on elements of data and text mining techniques. A simplified depiction of the process is provided in Figure 7.3.

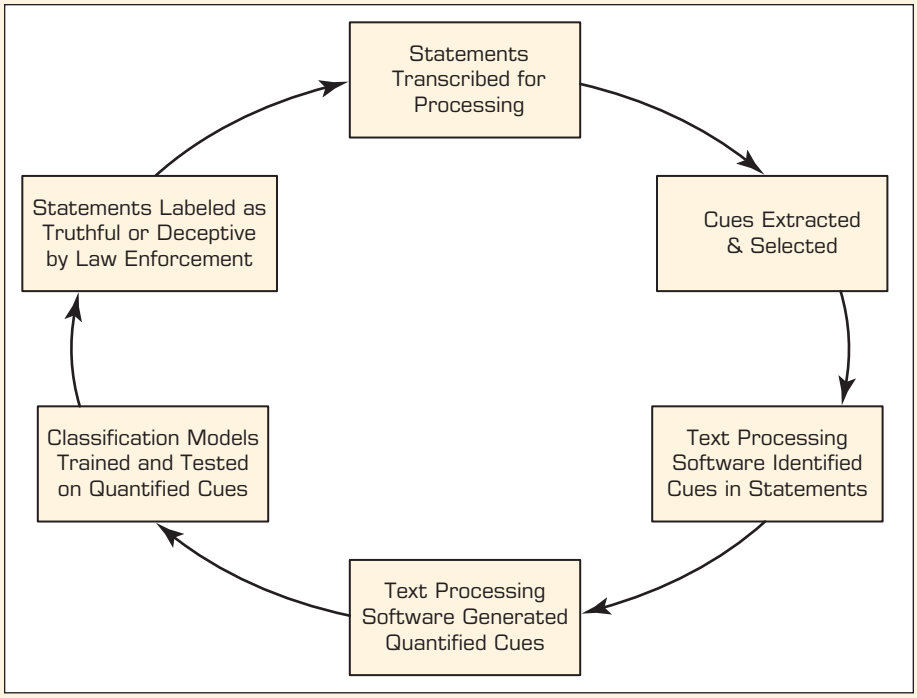


FIGURE 7.3 **Text-Based Deception-Detection Process.** *Source:* Fuller, C. M., D. Biros, & D. Delen. (2008, January). Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection. *Proceedings of the Forty-First Annual Hawaii International Conference on System Sciences (HICSS),* Big Island, HI: IEEE Press, pp. 80–99.

First, the researchers prepared the data for processing. The original handwritten statements had to be transcribed into a word processing file. Second, features (i.e., cues) were identified. The researchers identified 31 features representing categories or types of language that are relatively independent of the text content and that can be readily analyzed by automated means. For example, first-person pronouns such as *I* or *me* can be identified without analysis of the surrounding text. Table 7.1 lists the categories and examples of features used in this study.

TABLE 7.1 **Categories and Examples of Linguistic Features Used in Deception Detection**

Number	Construct (Category)	Example Cues
1	Quantity	Verb count, noun phrase count, etc.
2	Complexity	Average number of clauses, average sentence length, etc.
3	Uncertainty	Modifiers, modal verbs, etc.
4	Nonimmediacy	Passive voice, objectification, etc.
5	Expressivity	Emotiveness
6	Diversity	Lexical diversity, redundancy, etc.
7	Informality	Typographical error ratio
8	Specificity	Spatiotemporal information, perceptual information, etc.
9	Affect	Positive affect, negative affect, etc.

(*Continued*)

Application Case 7.3 (Continued)

The features were extracted from the textual statements and input into a flat file for further processing. Using several feature-selection methods along with 10-fold cross-validation, the researchers compared the prediction accuracy of three popular data mining methods. Their results indicated that neural network models performed the best, with 73.46 percent prediction accuracy on test data samples; decision trees performed second best, with 71.60 percent accuracy; and logistic regression was last, with 65.28 percent accuracy.

The results indicate that automated text-based deception detection has the potential to aid those who must try to detect lies in text and can be successfully applied to real-world data. The accuracy of these techniques exceeded the accuracy of most other deception-detection techniques, even though it was limited to textual cues.

Questions for Case 7.3

1. Why is it difficult to detect deception?
2. How can text/data mining be used to detect deception in text?
3. What do you think are the main challenges for such an automated system?

Sources: Fuller, C. M., D. Biros, & D. Delen. (2008, January). "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection." *Proceedings of the Forty-First Annual Hawaii International Conference on System Sciences (HICSS),* Big Island, HI: IEEE Press, pp. 80–99; Bond, C. F., & B. M. DePaulo. (2006). "Accuracy of Deception Judgments." *Personality and Social Psychology Reports, 10*(3), pp. 214–234.

this literature is relatively constant, having a fairly standardized ontology. What follows are a few exemplary studies that successfully used text mining techniques in extracting novel patterns from biomedical literature.

Experimental techniques such as DNA microarray analysis, serial analysis of gene expression (SAGE), and mass spectrometry proteomics, among others, are generating large amounts of data related to genes and proteins. As in any other experimental approach, it is necessary to analyze this vast amount of data in the context of previously known information about the biological entities under study. The literature is a particularly valuable source of information for experiment validation and interpretation. Therefore, the development of automated text mining tools to assist in such interpretation is one of the main challenges in current bioinformatics research.

Knowing the location of a protein within a cell can help to elucidate its role in biological processes and to determine its potential as a drug target. Numerous location-prediction systems are described in the literature; some focus on specific organisms, whereas others attempt to analyze a wide range of organisms. Shatkay et al. (2007) proposed a comprehensive system that uses several types of sequence- and text-based features to predict the location of proteins. The main novelty of their system lies in the way in which it selects its text sources and features and integrates them with sequence-based features. They tested the system on previously used and new data sets devised specifically to test its predictive power. The results showed that their system consistently beat previously reported results.

Chun et al. (2006) described a system that extracts disease–gene relationships from literature accessed via MEDLINE. They constructed a dictionary for disease and gene names from six public databases and extracted relation candidates by dictionary matching. Because dictionary matching produces a large number of false positives, they developed a method of machine-learning–based system, named entity recognition (NER), to filter out false recognition of disease/gene names. They found that the success of relation extraction is heavily dependent on the performance of NER filtering and that the filtering improved the precision of relation extraction by 26.7 percent at the cost of a small reduction in recall.

Figure 7.4 shows a simplified depiction of a multilevel text analysis process for discovering gene–protein relationships (or protein–protein interactions) in the biomedical

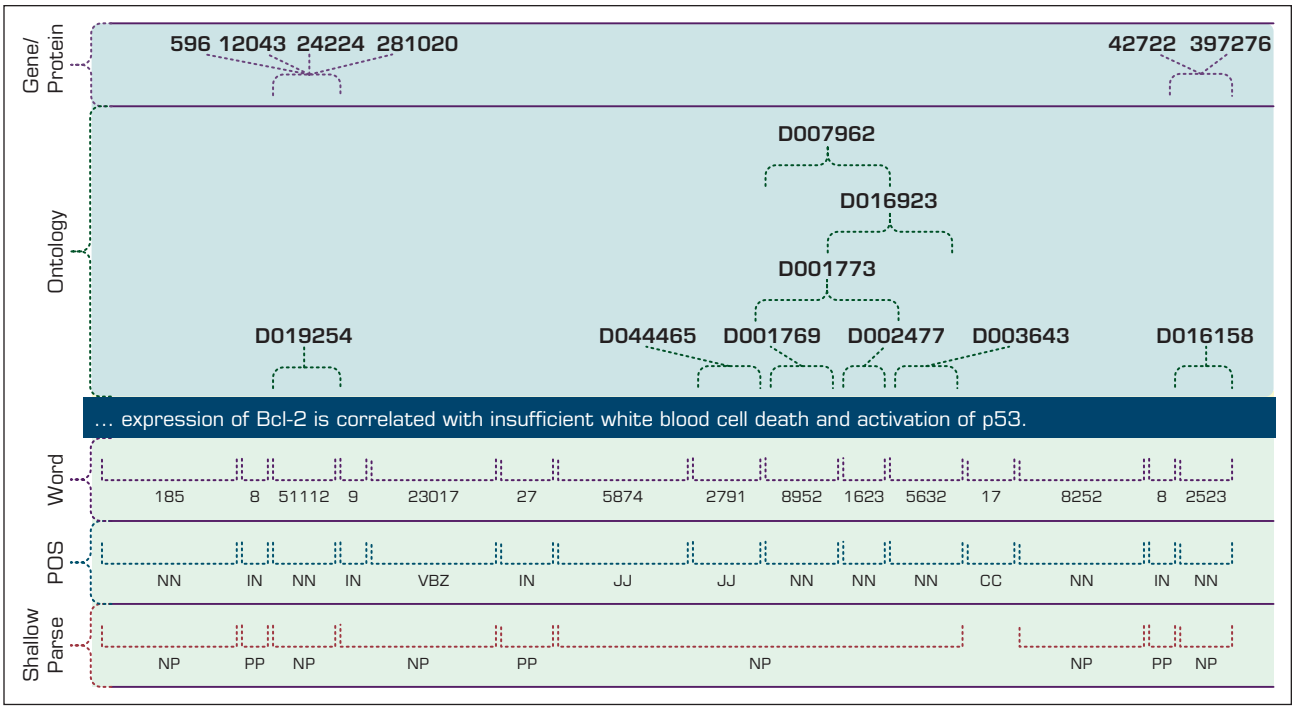


FIGURE 7.4 **Multilevel Analysis of Text for Gene/Protein Interaction Identification.** *Source:* Used with permission of Nakov, P., Schwartz, A., Wolf, B., & Hearst, M. A. (2005). Supporting annotation layers for natural language processing. *Proceedings of the Association for Computational Linguistics (ACL),* Interactive Poster and Demonstration Sessions, Ann Arbor, MI. Association for Computational Linguistics, 65–68.

literature (Nakov, Schwartz, Wolf, and Hearst, 2005). As can be seen in this simplified example that uses a simple sentence from biomedical text, first (at the bottom three levels) the text is tokenized using **part-of-speech tagging** and shallow parsing. The tokenized terms (words) are then matched (and interpreted) against the hierarchical representation of the domain ontology to derive the gene–protein relationship. Application of this method (and/or some variation of it) to the biomedical literature offers great potential to decode the complexities in the Human Genome Project.

Academic Applications

The issue of text mining is of great importance to publishers who hold large databases of information requiring indexing for better retrieval. This is particularly true in scientific disciplines in which highly specific information is often contained within written text. Initiatives have been launched, such as *Nature*'s proposal for an Open Text Mining Interface and the National Institutes of Health's common Journal Publishing Document Type Definition, which would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

Academic institutions have also launched text mining initiatives. For example, the National Centre for Text Mining, a collaborative effort between the Universities of Manchester and Liverpool, provides customized tools, research facilities, and advice on text mining to the academic community. With an initial focus on text mining in the biological and biomedical sciences, research has since expanded into the social sciences. In the United States, the School of Information at the University of California–Berkeley is developing a program called BioText to assist bioscience researchers in text mining and analysis.

As described in this section, text mining has a wide variety of applications in a number of different disciplines. See Application Case 7.4 for an example of how a leading computing product manufacturer uses text mining to better understand its current and potential customers' needs and wants related to product quality and product design.

Advanced analytics techniques that use both structured and unstructured data have been successfully used in many application domains. Application Case 7.4 provides an interesting example where a wide range of analytics capabilities are used to successfully manage the Orlando Magic organization both on and off the courts of NBA.

Application Case 7.4

The Magic Behind the Magic: Instant Access to Information Helps the Orlando Magic Up their Game and the Fan's Experience

From ticket sales to starting lineups, the Orlando Magic have come a long way since their inaugural season in 1989. There weren't many wins in those early years, but the franchise has weathered the ups and downs to compete at the highest levels of the NBA.

Professional sports teams in smaller markets often struggle to build a big enough revenue base to compete against their larger market rivals. By using SAS® Analytics and SAS® Data Management, the Orlando Magic are among the top revenue earners in the NBA, despite being in the 20th-largest market.

The Magic accomplish this feat by studying the resale ticket market to price tickets better, to predict season ticket holders at risk of defection (and lure them back), and to analyze concession and product merchandise sales to make sure the organization has what the fans want every time they enter the arena. The club has even used SAS to help coaches put together the best lineup.

"Our biggest challenge is to customize the fan experience, and SAS helps us manage all that in a robust way," says Alex Martins, CEO of the Orlando Magic. Having been with the Magic since the beginning (working his way up from PR Director to President to CEO), Martins has seen it all and knows the value that analytics adds. Under Martins' leadership, the season-ticket base has grown as large as 14,200, and the corporate sales department has seen tremendous growth.

The Challenge: Filling Every Seat

But like all professional sports teams, the Magic are constantly looking for new strategies that will keep the seats filled at each of the 41 yearly home games. "Generating new revenue streams in this day of escalating player salaries and escalating expenses is important," says Anthony Perez, vice president of Business Strategy. But with the advent of a robust online secondary market for tickets, reaching the industry benchmark of 90 percent renewal of season tickets has become more difficult.

"In the first year, we saw ticket revenue increase around 50 percent. Over the last three years—for that period, we've seen it grow maybe 75 percent. It's had a huge impact" said Anthony Perez, vice president of Business Strategy, Orlando Magic.

Perez's group takes a holistic approach by combining data from all revenue streams (concession, merchandise, and ticket sales) with outside data (secondary ticket market) to develop models that benefit the whole enterprise. "We're like an in-house consulting group," explains Perez.

In the case of season ticket holders, the team uses historical purchasing data and renewal patterns to build decision tree models that place subscribers into three categories: most likely to renew, least likely to renew, and fence sitters. The fence sitters then get the customer service department's attention come renewal time.

"SAS has helped us grow our business. It is probably one of the greatest investments that we've made as an organization over the last half-dozen years because we can point to top-line revenue growth that SAS has helped us create through the specific messaging that we're able to direct to each one of our client groups."

How Do They Predict Season Ticket Renewals?

When analytics showed the team that 80 percent of revenue was from season ticket holders, it decided to take a proactive approach to renewals and at-risk accounts. The Magic don't have a crystal ball, but they do have SAS® Enterprise Miner™, which allowed them to better understand their data and

develop analytic models that combine three pillars for predicting season ticket holder renewals:

- Tenure (how long had the customer been a ticket holder?).
- Ticket use (did the customer actually attend the games?).
- Secondary market activity (were the unused tickets successfully sold on secondary sites?).

The data mining tools allowed the team to accomplish more accurate scoring that led to a difference—and marked improvement—in the way it approached customer retention and marketing.

Ease of Use Helps Spread Analytics Message

Perez likes how easy it is to use SAS—it was a factor in opting to do the work in-house rather than outsourcing it. Perez's team has set up recurring processes and automated them. Data manipulation is minimal, "allowing us more time to interpret rather than just manually crunching the numbers." Business users throughout the organization, including executives, have instant access to information through SAS® Visual Analytics. "It's not just that we're using the tools daily; we are using them throughout the day to make decisions," Perez says.

Being Data-Driven

"We adopted an analytics approach years ago, and we're seeing it transform our entire organization," says Martins. "Analytics helps us understand customers better, helps in business planning (ticket pricing, etc.), and provides game-to-game and year-to-year data on demand by game and even by seat."

"And analytics has helped transform the game. GMs and analytics teams look at every aspect of the game, including movements of players on the court, to transform data to predict defense against certain teams. We can now ask ourselves, 'What are the most efficient lineups in a game? Which team can produce more points vs. another lineup? Which team is better defensively than another?'"

"We used to produce a series of reports manually, but now we can do it with five clicks of a mouse (instead of five hours overnight in anticipation of tomorrow's game). We can have dozens of reports available to staff in minutes. Analytics has made us smarter," says Martins.

What's Next?

"Getting real-time data is the next step for us in our analytical growth process," says Martins. "On a game day, getting real-time data to track what tickets are available and how to maximize yield of those tickets is critical. Additionally, you're going to see major technological changes and acceptance of the technology on the bench to see how the games are played moving forward. Maybe as soon as next season you'll see our assistant coaches with iPad® tablets getting real-time data, learning what the opponent is doing and what plays are working. It'll be necessary in the future.

"We're setting ourselves up to be successful moving forward. And in the very near future, we'll be in a position again to compete for a conference championship and an NBA championship," says Martins. "All of the moves made this year and the ones to come in the future will be done in order to build success on [and off] the court.''

Questions for Case 7.4

1. According to the application case, what were the main challenges the Orlando Magic was facing?
2. How did analytics help the Orlando Magic to overcome some of its most significant challenges on and off the court?
3. Can you think of other uses of analytics in sports and especially in the case of the Orlando Magic? You can search the Web to find some answers to this question.

Source: SAS Customer Story, "The magic behind the Magic: Instant access to information helps the Orlando Magic up their game and the fan's experience" at **https://www.sas.com/en_us/customers/orlando-magic.html** and **https://www.nba.com/magic/news/denton-25-years-magic-history** (accessed November 2018).

SECTION 7.4 REVIEW QUESTIONS

1. List and briefly discuss some of the text mining applications in marketing.
2. How can text mining be used in security and counterterrorism?
3. What are some promising text mining applications in biomedicine?

7.5 TEXT MINING PROCESS

To be successful, text mining studies should follow a sound methodology based on best practices. A standardized process model is needed similar to Cross-Industry Standard Process for Data Mining (CRISP-DM), which is the industry standard for data mining projects (see Chapter 4). Even though most parts of CRISP-DM are also applicable to text mining projects, a specific process model for text mining would include much more elaborate data preprocessing activities. Figure 7.5 depicts a high-level context diagram of a typical text mining process (Delen & Crossland, 2008). This context diagram presents the scope of the process, emphasizing its interfaces with the larger environment. In essence, it draws boundaries around the specific process to explicitly identify what is included in (and excluded from) the text mining process.

As the context diagram indicates, the input (inward connection to the left edge of the box) into the text-based knowledge-discovery process is the unstructured as well as structured data collected, stored, and made available to the process. The output (outward extension from the right edge of the box) of the process is the context-specific knowledge that can be used for decision making. The controls, also called the *constraints* (inward connection to the top edge of the box), of the process include software and hardware limitations, privacy issues, and difficulties related to processing the text that is presented in the form of natural language. The mechanisms (inward connection to the bottom edge of the box) of the process include proper techniques, software tools, and domain expertise. The primary purpose of text mining (within the context of knowledge discovery) is to process unstructured (textual) data (along with structured data if relevant to the problem being addressed and available) to extract meaningful and actionable patterns for better decision making.

At a very high level, the text mining process can be broken down into three consecutive tasks, each of which has specific inputs to generate certain outputs (see Figure 7.6). If, for some reason, the output of a task is not what is expected, a backward redirection to the previous task execution is necessary.

Task 1: Establish the Corpus

The main purpose of the first task activity is to collect all the documents related to the context (domain of interest) being studied. This collection may include textual documents, XML files, e-mails, Web pages, and short notes. In addition to the readily available

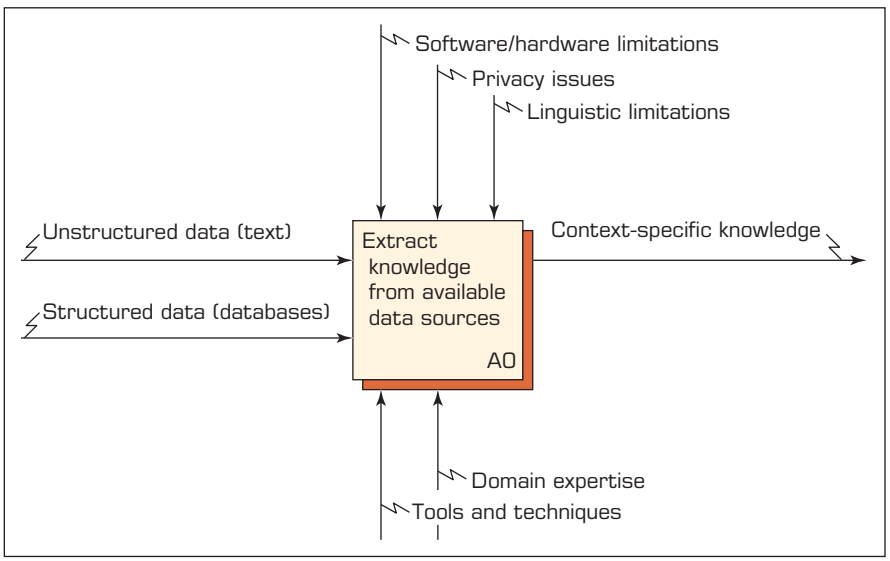


FIGURE 7.5 Context Diagram for the Text Mining Process.

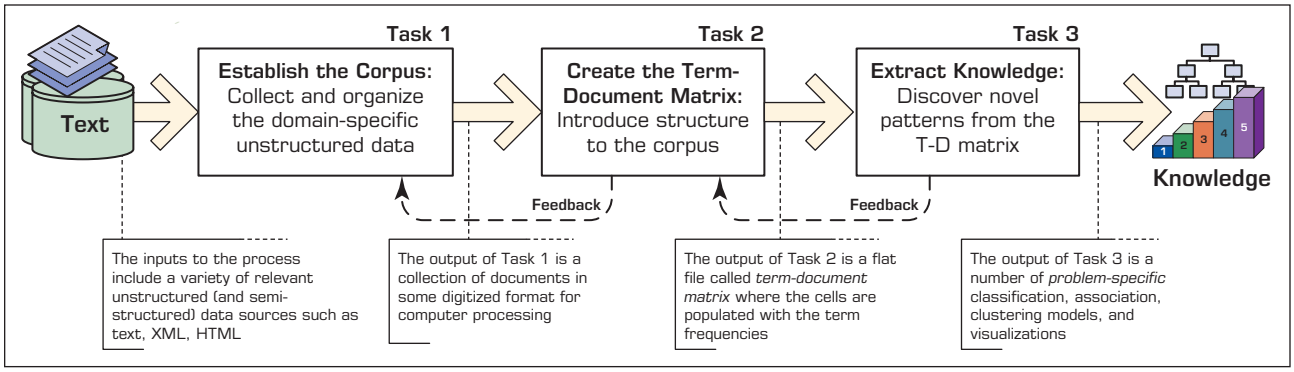


FIGURE 7.6 **The Three-Step/Task Text Mining Process.**

textual data, voice recordings may also be transcribed using speech-recognition algorithms and made a part of the text collection.

Once collected, the text documents are transformed and organized in a manner such that they are all in the same representational form (e.g., ASCII text files) for computer processing. The organization of the documents can be as simple as a collection of digitized text excerpts stored in a file folder or a list of links to a collection of Web pages in a specific domain. Many commercially available text mining software tools could accept these as input and convert them into a flat file for processing. Alternatively, the flat file can be prepared outside the text mining software and then presented as the input to the text mining application.

Task 2: Create the Term–Document Matrix

In this task, the digitized and organized documents (the corpus) are used to create the **term–document matrix (TDM)**. In the TDM, the rows represent the documents and the columns represent the terms. The relationships between the terms and documents are characterized by indices (i.e., a relational measure that can be as simple as the number of occurrences of the term in respective documents). Figure 7.7 is a typical example of a TDM.

Documents \ Terms	Investment Risk	Project Management	Software Engineering	Development	SAP	...	
Document 1	1			1			
Document 2		1					
Document 3			3		1		
Document 4		1					
Document 5			2	1			
Document 6	1			1			
...							

FIGURE 7.7 **Simple Term–Document Matrix.**

The goal is to convert the list of organized documents (the corpus) into a TDM where the cells are filled with the most appropriate indices. The assumption is that the essence of a document can be represented with a list and frequency of the terms used in that document. However, are all terms important when characterizing documents? Obviously, the answer is "no." Some terms, such as articles, auxiliary verbs, and terms used in almost all the documents in the corpus, have no differentiating power and, therefore, should be excluded from the indexing process. This list of terms, commonly called *stop terms* or *stop words*, is specific to the domain of study and should be identified by the domain experts. On the other hand, one might choose a set of predetermined terms under which the documents are to be indexed (this list of terms is conveniently called *include terms* or *dictionary*). In addition, synonyms (pairs of terms that are to be treated the same) and specific phrases (e.g., "Eiffel Tower") can also be provided so that the index entries are more accurate.

Another filtration that should take place to accurately create the indices is *stemming*, which refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of a verb are identified and indexed as the same word. For example, stemming will ensure that *modeling* and *modeled* will be recognized as the word *model*.

The first generation of the TDM includes all the unique terms identified in the corpus (as its columns), excluding the ones in the stop term list; all the documents (as its rows); and the occurrence count of each term for each document (as its cell values). If, as is commonly the case, the corpus includes a rather large number of documents, then there is a very good chance that the TDM will have a very large number of terms. Processing such a large matrix might be time consuming and, more important, might lead to extraction of inaccurate patterns. At this point, one has to decide the following: (1) What is the best representation of the indices? and (2) How can we reduce the dimensionality of this matrix to a manageable size?

REPRESENTING THE INDICES Once the input documents have been indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the extracted information. The raw term frequencies generally reflect on how salient or important a word is in each document. Specifically, words that occur with greater frequency in a document are better descriptors of the contents of that document. However, it is not reasonable to assume that the word counts themselves are proportional to their importance as descriptors of the documents. For example, if a word occurs one time in document *A* but three times in document *B*, it is not necessarily reasonable to conclude that this word is three times as important a descriptor of document *B* as compared to document *A*. To have a more consistent TDM for further analysis, these raw indices need to be normalized. As opposed to showing the actual frequency counts, the numerical representation between terms and documents can be normalized using a number of alternative methods, such as log frequencies, binary frequencies, and inverse document frequencies.

REDUCING THE DIMENSIONALITY OF THE MATRIX Because the TDM is often very large and rather sparse (most of the cells filled with zeros), another important question is, "How do we reduce the dimensionality of this matrix to a manageable size?" Several options are available for managing the matrix size:

- A domain expert goes through the list of terms and eliminates those that do not make much sense for the context of the study (this is a manual, labor-intensive process).
- Eliminate terms with very few occurrences in very few documents.
- Transform the matrix using SVD.

Singular value decomposition (SVD), which is closely related to principal components analysis, reduces the overall dimensionality of the input matrix (number of input documents by number of extracted terms) to a lower-dimensional space when each consecutive dimension represents the largest degree of variability (between words and documents) possible (Manning and Schutze, 1999). Ideally, the analyst might identify the two or three most salient dimensions that account for most of the variability (differences) between the words and documents, thus identifying the latent semantic space that organizes the words and documents in the analysis. Once such dimensions are identified, the underlying "meaning" of what is contained (discussed or described) in the documents has been extracted.

Task 3: Extract the Knowledge

Using the well-structured TDM and potentially augmented with other structured data elements, novel patterns are extracted in the context of the specific problem being addressed. The main categories of knowledge extraction methods are classification, clustering, association, and trend analysis. A short description of these methods follows.

CLASSIFICATION Arguably the most common knowledge-discovery topic in analyzing complex data sources is the **classification** (or categorization) of certain objects. The task is to classify a given data instance into a predetermined set of categories (or classes). As it applies to the domain of text mining, the task is known as *text categorization* for a given set of categories (subjects, topics, or concepts) and a collection of text documents whose goal is to find the correct topic (subject or concept) for each document using models developed with a training data set that includes both the documents and actual document categories. Today, automated text classification is applied in a variety of contexts, including automatic or semi-automatic (interactive) indexing of text, spam filtering, Web page categorization under hierarchical catalogs, automatic generation of metadata, and detection of genre.

The two main approaches to text classification are knowledge engineering and machine learning (Feldman and Sanger, 2007). With the knowledge-engineering approach, an expert's knowledge about the categories is encoded into the system either declaratively or in the form of procedural classification rules. With the machine-learning approach, a general inductive process builds a classifier by learning from a set of reclassified examples. As the number of documents increases at an exponential rate and as knowledge experts become harder to come by, the popularity trend between the two is shifting toward the machine-learning approach.

CLUSTERING **Clustering** is an unsupervised process whereby objects are classified into "natural" groups called *clusters*. Compared to categorization that uses a collection of preclassified training examples to develop a model based on the descriptive features of the classes to classify a new unlabeled example, in clustering the problem is to group an unlabeled collection of objects (e.g., documents, customer comments, Web pages) into meaningful clusters without any prior knowledge.

Clustering is useful in a wide range of applications from document retrieval to enabling better Web content searches. In fact, one of the prominent applications of clustering is the analysis and navigation of very large text collections, such as Web pages. The basic underlying assumption is that relevant documents tend to be more similar to each other than to irrelevant ones. If this assumption holds, the clustering of documents based on the similarity of their content improves search effectiveness (Feldman and Sanger, 2007):

- **Improved search recall.** Because it is based on overall similarity as opposed to the presence of a single term, clustering can improve the recall of a query-based search in such a way that when a query matches a document, its whole cluster is returned.

- **Improved search precision.** Clustering can also improve search precision. As the number of documents in a collection grows, it becomes difficult to browse through the list of matched documents. Clustering can help by grouping the documents into a number of much smaller groups of related documents, ordering them by relevance, and returning only the documents from the most relevant group (or groups).

The two most popular clustering methods are scatter/gather clustering and query-specific clustering:

- **Scatter/gather.** This document browsing method uses clustering to enhance the efficiency of human browsing of documents when a specific search query cannot be formulated. In a sense, the method dynamically generates a table of contents for the collection and adapts and modifies it in response to the user selection.
- **Query-specific clustering.** This method employs a hierarchical clustering approach where the most relevant documents to the posed query appear in small tight clusters that are nested in larger clusters containing less-similar documents, creating a spectrum of relevance levels among the documents. This method performs consistently well for document collections of realistically large sizes.

ASSOCIATION **Associations**, or *association rule learning in data mining*, is a popular and well-researched technique for discovering interesting relationships among variables in large databases. The main idea in generating association rules (or solving market-basket problems) is to identify the frequent sets that go together.

In text mining, associations specifically refer to the direct relationships between concepts (terms) or sets of concepts. The concept set association rule $A + C$ relating two frequent concept sets A and C can be quantified by the two basic measures of support and confidence. In this case, confidence is the percentage of documents that include all concepts in C within the same subset of those documents that include all concepts in A. Support is the percentage (or number) of documents that include all the concepts in A and C. For instance, in a document collection the concept "Software Implementation Failure" could appear most often in association with "Enterprise Resource Planning" and "Customer Relationship Management" with significant support (4%) and confidence (55%), meaning that 4 percent of the documents had all three concepts represented in the same document, and of the documents that included "Software Implementation Failure," 55 percent of them also included "Enterprise Resource Planning" and "Customer Relationship Management."

Text mining with association rules was used to analyze published literature (news and academic articles posted on the Web) to chart the outbreak and progress of the bird flu (Mahgoub et al., 2008). The idea was to automatically identify the association among the geographic areas, spreading across species, and countermeasures (treatments).

TREND ANALYSIS Recent methods of trend analysis in text mining have been based on the notion that the various types of concept distributions are functions of document collections; that is, different collections lead to different concept distributions for the same set of concepts. It is, therefore, possible to compare two distributions that are otherwise identical except that they are from different subcollections. One notable direction of this type of analysis is having two collections from the same source (such as from the same set of academic journals) but from different points in time. Delen and Crossland (2008) applied **trend analysis** to a large number of academic articles (published in the three highest-rated academic journals) to identify the evolution of key concepts in the field of information systems.

As described in this section, a number of methods are available for text mining. Application Case 7.5 describes the use of a number of different techniques in analyzing a large set of literature.

Application Case 7.5 Research Literature Survey with Text Mining

Researchers conducting searches and reviews of relevant literature face an increasingly complex and voluminous task. In extending the body of relevant knowledge, it has always been important to work hard to gather, organize, analyze, and assimilate existing information from the literature, particularly from one's home discipline. With the increasing abundance of potentially significant research being reported in related fields, and even in what are traditionally deemed to be nonrelated fields of study, the researcher's task is ever more daunting if a thorough job is desired.

In new streams of research, the researcher's task can be even more tedious and complex. Trying to ferret out relevant work that others have reported can be difficult, at best, and perhaps even nearly impossible if traditional, largely manual reviews of published literature are required. Even with a legion of dedicated graduate students or helpful colleagues, trying to cover all potentially relevant published work is problematic.

Many scholarly conferences take place every year. In addition to extending the body of knowledge of the current focus of a conference, organizers often desire to offer additional minitracks and workshops. In many cases, these additional events are intended to introduce attendees to significant streams of research in related fields of study and to try to identify the "next big thing" in terms of research interests and focus. Identifying reasonable candidate topics for such minitracks and workshops is often subjective rather than derived objectively from the existing and emerging research.

In a recent study, Delen and Crossland (2008) proposed a method to greatly assist and enhance the efforts of the researchers by enabling a semi-automated analysis of large volumes of published literature through the application of text mining. Using standard digital libraries and online publication search engines, the authors downloaded and collected all the available articles for the three major journals in the field of management information systems: *MIS Quarterly* (MISQ), *Information Systems Research* (ISR), and the *Journal of Management Information Systems* (JMIS). To maintain the same time interval for all three journals (for potential comparative longitudinal studies), the journal with the most recent starting date for its digital publication availability was used as the start time for this study (i.e., JMIS articles have been digitally available since 1994). For each article, Delen and Crossland extracted the title, abstract, author list, published keywords, volume, issue number, and year of publication. They then loaded all the article data into a simple database file. Also included in the combined data set was a field that designated the journal type of each article for likely discriminatory analysis. Editorial notes, research notes, and executive overviews were omitted from the collection. Table 7.2 shows how the data were presented in a tabular format.

In the analysis phase, the researchers chose to use only the abstract of an article as the source of information extraction. They chose not to include the keywords listed with the publications for two main reasons: (1) under normal circumstances, the abstract would already include the listed keywords, and therefore inclusion of the listed keywords for the analysis would mean repeating the same information and potentially giving them unmerited weight and (2) the listed keywords could be terms that authors would like their article to be associated with (as opposed to what is really contained in the article), therefore, potentially introducing unquantifiable bias to the analysis of the content.

The first exploratory study was to look at the longitudinal perspective of the three journals (i.e., evolution of research topics over time). To conduct a longitudinal study, Delen and Crossland divided the 12-year period (from 1994 to 2005) into four 3-year periods for each of the three journals. This framework led to 12 text mining experiments with 12 mutually exclusive data sets. At this point, for each of the 12 data sets, the researchers used text mining to extract the most descriptive terms from these collections of articles represented by their abstracts. The results were tabulated and examined for time-varying changes in the terms published in these three journals.

(*Continued*)

Application Case 7.5 (Continued)

TABLE 7.2 Tabular Representation of the Fields Included in the Combined Data Set

A1 fx ID

	A	B	C	D	E	F	G	H	I	J
1	ID	YEAR	JOURNAL	ABSTRACT						
2	PID001	2005	MISQ	The need for continual value innovation is driving supply chains to evolve from						
3	PID002	1999	ISR	Although much contemporary thought considers advanced information technc						
4	PID003	2001	JMIS	When producers of goods (or services) are confronted by a situation in which						
5	PID004	1995	ISR	Preservation of organizational memory becomes increasingly important to org						
6	PID005	1994	ISR	The research reported here is an adaptation of a model developed to measure						
7	PID006	1995	MISQ	This study evaluates the extent to which the added value to customers from a						
8	PID007	2003	MISQ	This paper reports the results(-) of a field-study of six medical project teams t						
9	PID008	1999	JMIS	Researchers and managers are beginning to realize that the full advantages c						
10	PID009	2000	JMIS	The Internet commerce technologies have significantly reduced sellers' costs						
11	PID010	1997	ISR	Adaptive Structuration Theory (AST) is rapidly becoming an influential theoret						
12	PID011	1995	JMIS	Research shows that group support systems (GSS) have dramatically increa						
13	PID012	2000	MISQ	Increasingly, business leaders are demanding that IT play the role of a busine						
14	PID013	2001	ISR	Alignment between business strategy and IS strategy is widely believed to im						
15	PID014	1999	JMIS	A framework is outlined that includes the planning of and setting goals for IT,						
16	PID015	1999	JMIS	The continuously growing importance of information technology (IT) requires c						
17	PID016	1994	MISQ	Identifying the best way to organize the IS functions within an interprise has b						
18	PID017	1996	ISR	Reasons for the mixed reactions to todays electronic off-exchange trading sy						
19	PID018	1996	JMIS	The performance impacts of information technology investments in organizati						
20	PID019	1997	JMIS	Anonymity is a fundamental concept in group support systems (GSS) resear						
21	PID020	2002	ISR	Although electronic commerce (EC) has created new opportunities for busine						
22	PID021	2005	JMIS	Understanding the successful adoption of information technology is largely ba						
23	PID022	2005	MISQ	Enterprise resource planning (ERP) systems and other complex information s						
24	PID023	1994	JMIS	Model management systems support modelers in various phases of the mod						
25	PID024	1995	ISR	While computer training is widely recognized as an essential contributor to th						

As a second exploration, using the complete data set (including all three journals and all four periods), Delen and Crossland conducted a clustering analysis. Clustering is arguably the most commonly used text mining technique. Clustering was used in this study to identify the natural groupings of the articles (by putting them into separate clusters) and then to list the most descriptive terms that characterized those clusters. They used SVD to reduce the dimensionality of the term-by-document matrix and then an expectation-maximization algorithm to create the clusters. They conducted several experiments to identify the *optimal* number of clusters, which turned out to be nine. After the construction of the nine clusters, they analyzed the content of those clusters from two perspectives: (1) representation of the journal type (see Figure 7.8a) and (2) representation of time (Figure 7.8b). The idea was to explore the potential differences and/or commonalities among the three journals and potential changes in the emphasis on those clusters; that is, to answer questions such as "Are there clusters that represent different research themes specific to a single journal?" and "Is there a time-varying characterization of those clusters?" The researchers discovered and discussed several interesting patterns using tabular and graphical representation of their findings (for further information, see Delen and Crossland, 2008).

Questions for Case 7.5

1. How can text mining be used to ease the insurmountable task of literature review?
2. What are the common outcomes of a text mining project on a specific collection of journal articles? Can you think of other potential outcomes not mentioned in this case?

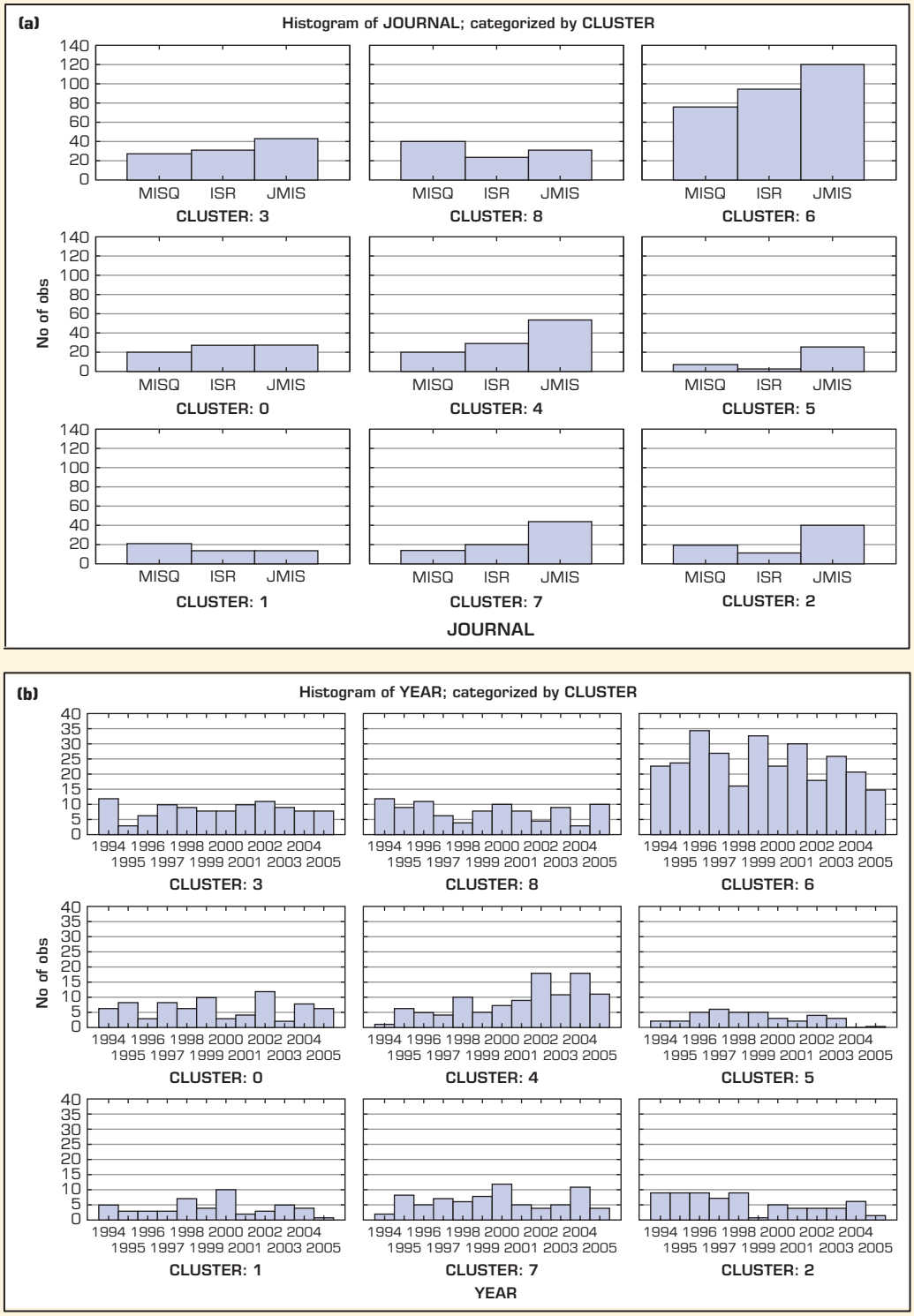


FIGURE 7.8 **(a) Distribution of the Number of Articles for the Three Journals over the Nine Clusters. (b) Development of the Nine Clusters over the Years.**

Source: Used with permission of Delen, D., & M. Crossland. (2008). "Seeding the Survey and Analysis of Research Literature with Text Mining." *Expert Systems with Applications, 34*(3), pp. 1707–1720.

SECTION 7.5 REVIEW QUESTIONS

1. What are the main steps in the text mining process?
2. What is the reason for normalizing word frequencies? What are the common methods for normalizing word frequencies?
3. What is SVD? How is it used in text mining?
4. What are the main knowledge extraction methods from corpus?

7.6 SENTIMENT ANALYSIS

We humans are social beings. We are adept at utilizing a variety of means to communicate. We often consult financial discussion forums before making an investment decision; ask our friends for their opinions on a newly opened restaurant or a newly released movie; and conduct Internet searches and read consumer reviews and expert reports before making a big purchase like a house, a car, or an appliance. We rely on others' opinions to make better decisions, especially in an area where we do not have much knowledge or experience. Thanks to the growing availability and popularity of opinion-rich Internet resources such as social media outlets (e.g., Twitter, Facebook), online review sites, and personal blogs, it is now easier than ever to find opinions of others (thousands of them, as a matter of fact) on everything from the latest gadgets to political and public figures. Even though not everybody expresses opinions over the Internet—due mostly to the fast-growing number and capabilities of social communication channels—the numbers are increasing exponentially.

Sentiment is a difficult word to define. It is often linked to or confused with other terms like *belief, view, opinion*, and *conviction*. Sentiment suggests a settled opinion reflective of one's feelings (Mejova, 2009). Sentiment has some unique properties that set it apart from other concepts that we might want to identify in text. Often we want to categorize text by topic, which could involve dealing with whole taxonomies of topics. Sentiment classification, on the other hand, usually deals with two classes (positive versus negative), a range of polarity (e.g., star ratings for movies), or even a range in strength of opinion (Pang and Lee, 2008). These classes span many topics, users, and documents. Although dealing with only a few classes might seem like an easier task than standard text analysis, this is far from the truth.

As a field of research, sentiment analysis is closely related to computational linguistics, NLP, and text mining. Sentiment analysis has many names. It is often referred to as *opinion mining, subjectivity analysis*, and *appraisal extraction* with some connections to affective computing (computer recognition and expression of emotion). The sudden upsurge of interest and activity in the area of sentiment analysis (i.e., opinion mining), which deals with the automatic extraction of opinions, feelings, and subjectivity in text, is creating opportunities and threats for businesses and individuals alike. The ones who embrace and take advantage of it will greatly benefit from it. Every opinion put on the Internet by an individual or a company will be accredited to the originator (good or bad) and will be retrieved and mined by others (often automatically by computer programs).

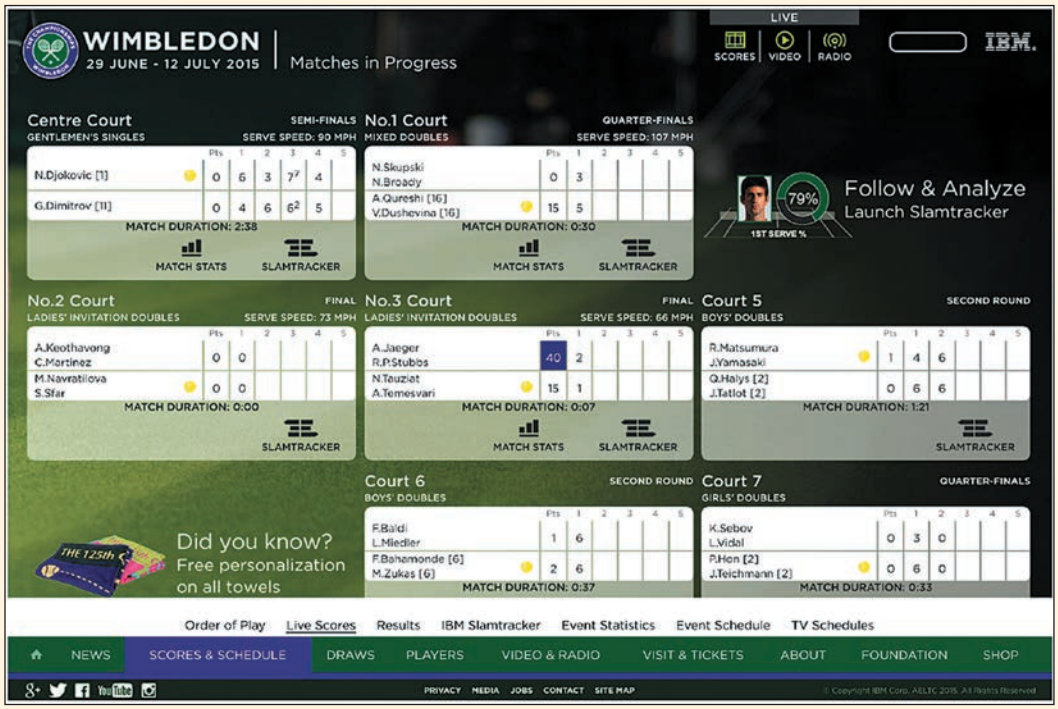
Sentiment analysis is trying to answer the question, "What do people feel about a certain topic?" by digging into opinions held by many using a variety of automated tools. Bringing together researchers and practitioners in business, computer science, computational linguistics, data mining, text mining, psychology, and even sociology, sentiment analysis aims to expand the traditional fact-based text analysis to new frontiers, to realize opinion-oriented information systems. In a business setting, especially in marketing and CRM, sentiment analysis seeks to detect favorable and unfavorable opinions toward specific products and/or services using large numbers of textual data sources (customer feedback in the form of Web postings, tweets, blogs, etc.).

Sentiment that appears in text comes in two flavors: explicit in which the subjective sentence directly expresses an opinion ("It's a wonderful day"), and implicit in which the text

implies an opinion ("The handle breaks too easily"). Most of the earlier work done in sentiment analysis focused on the first kind of sentiment because it is easier to analyze. Current trends are to implement analytical methods to consider both implicit and explicit sentiments. Sentiment polarity is a particular feature of text on which sentiment analysis primarily focuses. It is usually dichotomized into two—positive and negative—but polarity can also be thought of as a range. A document containing several opinionated statements will have a mixed polarity overall, which is different from not having a polarity at all (being objective; Mejova, 2009). Timely collection and analysis of textual data, which may be coming from a variety of sources—ranging from customer call center transcripts to social media postings—is a crucial part of the capabilities of proactive and customer-focused companies today. These real-time analyses of textual data are often visualized in easy-to-understand dashboards. Application Case 7.6 provides a customer success story in which a collection of analytics solutions is collectively used to enhance viewers' experience at the Wimbledon tennis tournament.

Application Case 7.6

Creating a Unique Digital Experience to Capture Moments That Matter at Wimbledon



Live scores displayed on Wimbledon.com, the official website of the championships, wimbledon, Copyright AELTC and IBM. Used with permission.

Known to millions of fans simply as "Wimbledon," The Championships are the oldest of tennis's four Grand Slams, and one of the world's highest-profile sporting events. Organized by the All England Lawn Tennis Club (AELTC), it has been a global sporting and cultural institution since 1877.

The Champion of Championships

The organizers of The Championships, Wimbledon, and AELTC have a simple objective: every year, they want to host the best tennis championships in the world—in every way and by every metric.

(*Continued*)

Application Case 7.6 (Continued)

The motivation behind this commitment is not simply pride; it also has a commercial basis. Wimbledon's brand is built on its premier status; this is what attracts both fans and partners. The world's best media organizations and greatest corporations—IBM included—want to be associated with Wimbledon precisely because of its reputation for excellence.

For this reason, maintaining the prestige of The Championships is one of AELTC's top priorities, but there are only two ways that the organization can directly control how the rest of the world perceives The Championships.

The first, and most important, is to provide an outstanding experience for the players, journalists, and spectators who are lucky enough to visit and watch the tennis courtside. AELTC has vast experience in this area. Since 1877, it has delivered two weeks of memorable, exciting competition in an idyllic setting: tennis in an English country garden.

The second is The Championships' online presence, which is delivered via the **wimbledon.com** Web site, mobile apps, and social media channels. The constant evolution of these digital platforms is the result of a 26-year partnership between AELTC and IBM.

Mick Desmond, commercial and media director at AELTC, explains, "When you watch Wimbledon on TV, you are seeing it through the broadcaster's lens. We do everything we can to help our media partners put on the best possible show, but at the end of the day, their broadcast is their presentation of The Championships."

He adds, "Digital is different: it's our platform, where we can speak directly to our fans—so it's vital that we give them the best possible experience. No sporting event or media channel has the right to demand a viewer's attention, so if we want to strengthen our brand, we need people to see our digital experience as the number-one place to follow The Championships online."

To that end, AELTC set a target of attracting 70 million visits, 20 million unique devices, and 8 million social followers during the two weeks of The Championships in 2015. It was up to IBM and AELTC to find a way to deliver.

Delivering a Unique Digital Experience

IBM and AELTC embarked on a complete redesign of the digital platform, using their intimate knowledge of The Championships' audience to develop an experience tailor-made to attract and retain tennis fans from across the globe.

"We recognized that while mobile is increasingly important, 80% of our visitors are using desktop computers to access our Web site," says Alexandra Willis, head of Digital and Content at AELTC. She continued,

> Our challenge for 2015 was how to update our digital properties to adapt to a mobile-first world, while still offering the best possible desktop experience. We wanted our new site to take maximum advantage of that large screen size and give desktop users the richest possible experience in terms of high-definition visuals and video content—while also reacting and adapting seamlessly to smaller tablet or mobile formats.
>
> Second, we placed a major emphasis on putting content in context—integrating articles with relevant photos, videos, stats and snippets of information, and simplifying the navigation so that users could move seamlessly to the content that interests them most.

On the mobile side, the team recognized that the wider availability of high bandwidth 4G connections meant that the mobile Web site would become more popular than ever—and ensured that it would offer easy access to all rich media content. At the same time, The Championships' mobile apps were enhanced with real-time notifications of match scores and events—and could even greet visitors as they passed through stations on the way to the grounds.

The team also built a special set of Web sites for the most important tennis fans of all: the players themselves. Using IBM® Bluemix® technology, it built a secure Web application that provided players a personalized view of their court bookings, transport, and on-court times, as well as helping them review their performance with access to stats on every match they played.

Turning Data into Insight—and Insight into Narrative

To supply its digital platforms with the most compelling possible content, the team took advantage of a unique opportunity: its access to real-time, shot-by-shot data on every match played during

The Championships. Over the course of the Wimbledon fortnight, 48 courtside experts capture approximately 3.4 million data points, tracking the type of shot, strategies, and outcome of each and every point.

These data are collected and analyzed in real time to produce statistics for TV commentators and journalists—and for the digital platform's own editorial team.

Willis went on to explain:

> This year IBM gave us an advantage that we had never had before—using data streaming technology to provide our editorial team with real-time insight into significant milestones and breaking news.
>
> The system automatically watched the streams of data coming in from all 19 courts, and whenever something significant happened—such as Sam Groth hitting the second-fastest serve in Championships' history—it let us know instantly. Within seconds, we were able to bring that news to our digital audience and share it on social media to drive even more traffic to our site.
>
> The ability to capture the moments that matter and uncover the compelling narratives within the data, faster than anyone else, was key. If you wanted to experience the emotions of The Championships live, the next best thing to being there in person was to follow the action on **wimbledon.com**.

Harnessing the Power of Natural Language

Another new capability tried in 2015 was the use of IBM's NLP technologies to help mine AELTC's huge library of tennis history for interesting contextual information. The team trained IBM Watson™ Engagement Advisor to digest this rich unstructured data set and use it to answer queries from the press desk.

The same NLP front-end was also connected to a comprehensive structured database of match statistics, dating back to the first Championships in 1877—providing a one-stop shop for both basic questions and more complex inquiries.

"The Watson trial showed a huge amount of potential. Next year, as part of our annual innovation planning process, we will look at how we can use it more widely—ultimately in pursuit of giving fans more access to this incredibly rich source of tennis knowledge," says Desmond.

Taking to the Cloud

IBM hosted the whole digital environment in its Hybrid Cloud. IBM used sophisticated modeling techniques to predict peaks in demand based on the schedule, popularity of each player, time of day, and many other factors—enabling it to dynamically allocate cloud resources appropriately to each piece of digital content and ensure a seamless experience for millions of visitors around the world.

In addition to the powerful private cloud platform that has supported The Championships for several years, IBM also used a separate SoftLayer® cloud to host the Wimbledon Social Command Centre and provide additional incremental capacity to supplement the main cloud environment during times of peak demand.

The elasticity of the cloud environment is key because The Championships' digital platforms need to be able to scale efficiently by a factor of more than 100 within a matter of days as the interest builds ahead of the first match on Centre Court.

Keeping Wimbledon Safe and Secure

Online security is a key concern today for all organizations. For major sporting events in particular, brand reputation is everything—and while the world is watching, it is particularly important to avoid becoming a high-profile victim of cyber crime. For these reasons, security has a vital role to play in IBM's partnership with AELTC.

Over the first five months of 2015, IBM security systems detected a 94 percent increase in security events on the **wimbledon.com** infrastructure compared to the same period in 2014.

As security threats—in particular distributed denial of service (DDoS) attacks—become ever more prevalent, IBM continually increases its focus on providing industry-leading levels of security for AELTC's entire digital platform.

A full suite of IBM security products, including IBM QRadar® SIEM and IBM Preventia Intrusion Prevention, enabled the 2015 Championships to run smoothly and securely and the digital platform to deliver a high-quality user experience at all times.

(*Continued*)

Application Case 7.6 (Continued)

Capturing Hearts and Minds

The success of the new digital platform for 2015—supported by IBM cloud, analytics, mobile, social, and security technologies—was immediate and complete. Targets for total visits and unique visitors were not only met but also exceeded. Achieving 71 million visits and 542 million page views from 21.1 million unique devices demonstrates the platform's success in attracting a larger audience than ever before and keeping those viewers engaged throughout The Championships.

"Overall, we had 13% more visits from 23% more devices than in 2014, and the growth in the use of **wimbledon.com** on mobile was even more impressive," says Willis. "We saw 125% growth in unique devices on mobile, 98% growth in total visits, and 79% growth in total page views."

Desmond concludes, "The results show that in 2015, we won the battle for fans' hearts and minds. People may have favorite newspapers and sports website that they visit for 50 weeks of the year—but for two weeks, they came to us instead."

He continued, "That's a testament to the sheer quality of the experience we can provide—harnessing our unique advantages to bring them closer to the action than any other media channel. The ability to capture and communicate relevant content in real time helped our fans experience The Championships more vividly than ever before."

Questions for Case 7.6

1. How did Wimbledon use analytics capabilities to enhance viewers' experience?
2. What were the challenges, proposed solution, and obtained results?

Source: IBM Case Study. "Creating a Unique Digital Experience to Capture the Moments That Matter." **http://www-03.ibm.com/software/businesscasestudies/us/en/corp?synkey=D140192K15783Q68** (accessed May 2016).

Sentiment Analysis Applications

Compared to traditional sentiment analysis methods, which were survey based or focus group centered, costly, and time consuming (and therefore driven from a small sample of participants), the new face of text analytics–based sentiment analysis is a limit breaker. Current solutions automate very large-scale data collection, filtering, classification, and clustering methods via NLP and data mining technologies that handle both factual and subjective information. Sentiment analysis is perhaps the most popular application of text analytics, tapping into data sources such as tweets, Facebook posts, online communities, discussion boards, Web logs, product reviews, call center logs and recordings, product rating sites, chat rooms, price comparison portals, search engine logs, and newsgroups. The following applications of sentiment analysis are meant to illustrate the power and the widespread coverage of this technology.

VOICE OF THE CUSTOMER **Voice of the customer (VOC)** is an integral part of analytic CRM and customer experience management systems. As the enabler of VOC, sentiment analysis can access a company's product and service reviews (either continuously or periodically) to better understand and better manage customer complaints and compliments. For instance, a motion picture advertising/marketing company can detect negative sentiments about a movie that is soon to open in theatres (based on its trailers) and quickly change the composition of trailers and advertising strategy (on all media outlets) to mitigate the negative impact. Similarly, a software company can detect the negative buzz regarding the bugs found in their newly released product early enough to release patches and quick fixes to alleviate the situation.

Often, the focus of VOC is individual customers, their service- and support-related needs, wants, and issues. VOC draws data from the full set of customer touch points, including e-mails, surveys, call center notes/recordings, and social media postings, and matches customer voices to transactions (inquiries, purchases, returns) and individual customer profiles captured in enterprise operational systems. VOC, mostly driven by sentiment analysis, is a key element of customer experience management initiatives, where the goal is to create an intimate relationship with the customer.

VOICE OF THE MARKET (VOM) VOM is about understanding aggregate opinions and trends. It is about knowing what stakeholders—customers, potential customers, influencers, whoever—are saying about your (and your competitors') products and services. A well-done VOM analysis helps companies with competitive intelligence and product development and positioning.

VOICE OF THE EMPLOYEE (VOE) Traditionally, VOE has been limited to employee satisfaction surveys. Text analytics in general (and sentiment analysis in particular) is a huge enabler of assessing the VOE. Using rich, opinionated textual data provides an effective and efficient way to listen to what employees are saying. As we all know, happy employees empower customer experience efforts and improve customer satisfaction.

BRAND MANAGEMENT Brand management focuses on listening to social media where anyone (past/current/prospective customers, industry experts, other authorities) can post opinions that can damage or boost a company's reputation. A number of relatively newly launched start-up companies offer analytics-driven brand management services for others. Brand management is product and company (rather than customer) focused. It attempts to shape perceptions rather than to manage experiences using sentiment analysis techniques.

FINANCIAL MARKETS Predicting the future values of individual (or a group of) stocks has been an interesting and seemingly unsolvable problem. What makes a stock (or a group of stocks) move up or down is anything but an exact science. Many believe that the stock market is mostly sentiment driven, making it anything but rational (especially for short-term stock movements). Therefore, the use of sentiment analysis in financial markets has gained significant popularity. Automated analysis of market sentiment using social media, news, blogs, and discussion groups seems to be a proper way to compute the market movements. If done correctly, sentiment analysis can identify short-term stock movements based on the buzz in the market, potentially impacting liquidity and trading.

POLITICS As we all know, opinions matter a great deal in politics. Because political discussions are dominated by quotes, sarcasm, and complex references to persons, organizations, and ideas, politics is one of the most difficult, and potentially fruitful, areas for sentiment analysis. By analyzing the sentiment on election forums, one might predict who is more likely to win or lose a race. Sentiment analysis can help understand what voters are thinking and can clarify a candidate's position on issues. Sentiment analysis can help political organizations, campaigns, and news analysts to better understand which issues and positions matter the most to voters. The technology was successfully applied by both parties to the 2008 and 2012 U.S. presidential election campaigns.

GOVERNMENT INTELLIGENCE Government intelligence is another application that has been used by intelligence agencies. For example, it has been suggested that one could monitor sources for increases in hostile or negative communications. Sentiment analysis can allow the automatic analysis of the opinions that people submit about pending policy

or government regulation proposals. Furthermore, monitoring communications for spikes in negative sentiment could be of use to agencies such as Homeland Security.

OTHER INTERESTING AREAS Sentiments of customers can be used to better design e-commerce sites (product suggestions, up-sell/cross-sell advertising), better place advertisements (e.g., placing dynamic advertisements of products and services that consider the sentiment on the page the user is browsing), and manage opinion- or review-oriented search engines (i.e., an opinion-aggregation Web site, an alternative to sites similar to Epinions, summarizing user reviews). Sentiment analysis can help with e-mail filtration by categorizing and prioritizing incoming e-mails (e.g., it can detect strongly negative or flaming e-mails and forward them to a proper folder), and citation analysis can determine whether an author is citing a piece of work as supporting evidence or in research but dismisses.

Sentiment Analysis Process

Because of the complexity of the problem (underlying concepts, expressions in text, context in which text is expressed, etc.), there is no readily available standardized process to conduct sentiment analysis. However, based on the published work in the field of sensitivity analysis so far (both on research methods and range of applications), a multistep, simple logical process as given in Figure 7.9 seems to be an appropriate methodology for sentiment analysis. These logical steps are iterative (i.e., feedback, corrections, and iterations are part of the discovery process) and experimental in nature, and once completed and combined, capable of producing desired insight about the opinions in the text collection.

STEP 1: SENTIMENT DETECTION After retrieval and preparation of the text documents, the first main task in sensitivity analysis is the detection of objectivity. Here the goal is to differentiate between a fact and an opinion, which can be viewed as classification of text as objective or subjective. This can also be characterized as calculation of Objectivity–Subjectivity ([O−S] Polarity, which can be represented with a numerical value ranging from 0 to 1). If the objectivity value is close to 1, there is no opinion to mine (i.e., it is a fact); therefore, the process goes back and grabs the next text data to analyze. Usually opinion detection is based on the examination of adjectives in text. For example, the polarity of "what a wonderful work" can be determined relatively easily by looking at the adjective.

STEP 2: N–P (NEGATIVE OR POSITIVE) POLARITY CLASSIFICATION The second main task is that of polarity classification. Given an opinionated piece of text, the goal is to classify the opinion as falling under one of two opposing sentiment polarities or to locate its position on the continuum between these two polarities (Pang & Lee, 2008). When viewed as a binary feature, polarity classification is the binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion (e.g., thumbs up or thumbs down). In addition to the identification of N–P polarity, one should also be interested in identifying the strength of the sentiment (as opposed to just positive, it can be expressed as mildly, moderately, strongly, or very strongly positive). Most of this research was done on product or movie reviews where the definitions of "positive" and "negative" are quite clear. Other tasks, such as classifying news as "good" or "bad," present some difficulty. For instance, an article could contain negative news without explicitly using any subjective words or terms. Furthermore, these classes usually appear intermixed when a document expresses both positive and negative sentiments. Then the task can be to identify the main (or dominating) sentiment of the document.

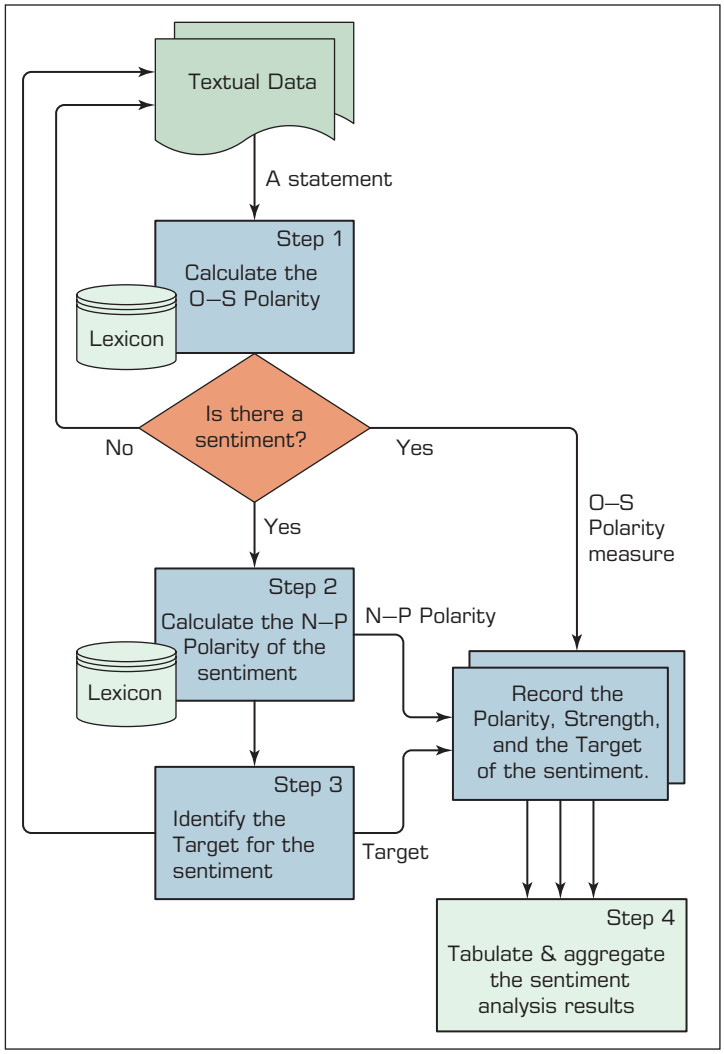


FIGURE 7.9 **Multistep Process to Sentiment Analysis.**

Still, for lengthy texts, the tasks of classification might need to be done at several levels: term, phrase, sentence, and perhaps document level. For those, it is common to use the outputs of one level as the inputs for the next higher layer. Several methods used to identify the polarity and strengths of the polarity are explained in the next section.

STEP 3: TARGET IDENTIFICATION The goal of this step is to accurately identify the target of the expressed sentiment (e.g., a person, a product, an event). The difficulty of this task depends largely on the domain of the analysis. Even though it is usually easy to accurately identify the target for product or movie reviews because the review is directly connected to the target, it can be quite challenging in other domains. For instance, lengthy, general-purpose text such as Web pages, news articles, and blogs do not always have a predefined topic assigned to them and often mention many objects, any of which could be deduced as the target. Sometimes there is more than one target in a sentiment sentence, which is the case in comparative texts. A subjective comparative sentence orders objects in order of preferences—for example, "This laptop computer is better than my desktop PC." These sentences can be identified using

comparative adjectives and adverbs (more, less, better, longer), superlative adjectives (most, least, best), and other words (such as same, differ, win, prefer). Once the sentences have been retrieved, the objects can be put in an order that is most representative of their merits as described in the text.

STEP 4: COLLECTION AND AGGREGATION Once the sentiments of all text data points in the document have been identified and calculated, in this step they are aggregated and converted to a single sentiment measure for the entire document. This aggregation could be as simple as summing up the polarities and strengths of all texts or as complex as using semantic aggregation techniques from NLP to identify the ultimate sentiment.

Methods for Polarity Identification

As mentioned in the previous section, **polarity identification** can be made at the word, term, sentence, or document level. The most granular level for polarity identification is at the word level. Once the polarity identification has been made at the word level, then it can be aggregated to the next higher level, and then the next until the level of aggregation desired from the sentiment analysis is reached. Two dominant techniques have been used for identification of polarity at the word/term level, each having its advantages and disadvantages:

1. Using a lexicon as a reference library (developed either manually or automatically by an individual for a specific task or developed by an institution for general use).
2. Using a collection of training documents as the source of knowledge about the polarity of terms within a specific domain (i.e., inducing predictive models from opinionated textual documents).

Using a Lexicon

A *lexicon* is essentially the catalog of words, their synonyms, and their meanings for a given language. In addition to lexicons for many other languages, there are several general-purpose lexicons created for English. Often general-purpose lexicons are used to create a variety of special-purpose lexicons for use in sentiment analysis projects. Perhaps the most popular general-purpose lexicon is WordNet created at Princeton University; it has been extended and used by many researchers and practitioners for sentiment analysis purposes. As described on the WordNet Web site (**wordnet.princeton.edu**), it is a large lexical database of English, including nouns, verbs, adjectives, and adverbs grouped into sets of cognitive synonyms (i.e., synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual–semantic and lexical relations.

An interesting extension of WordNet was created by Esuli and Sebastiani (2006) where they added polarity (Positive–Negative; P–N) and objectivity (Subjective–Objective; S–O) labels for each term in the lexicon. To label each term, they classified the synset (a group of synonyms) to which a term belongs using a set of ternary classifiers (a measure that attaches to each object exactly one of three labels), each capable of deciding whether a synset is positive, or negative, or objective. The resulting scores range from 0.0 to 1.0, giving a graded evaluation of opinion-related properties of the terms. These can be summed up visually as in Figure 7.10. The edges of the triangle represent one of the three classifications (positive, negative, and objective). A term can be located in this space as a point representing the extent to which it belongs to each of the classifications.

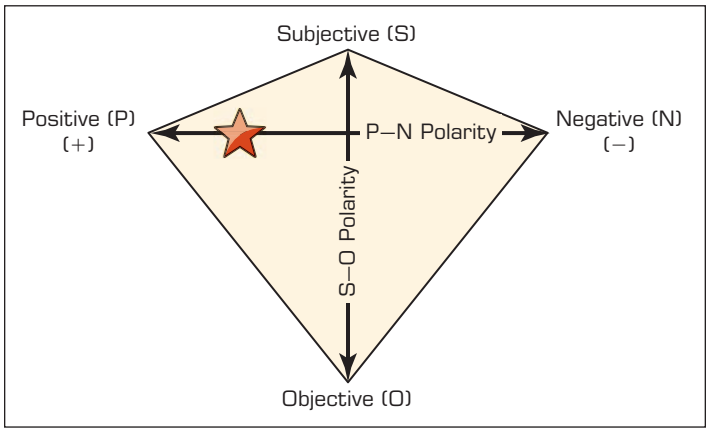


FIGURE 7.10 **Graphical Representation of the P–N Polarity and S–O Polarity Relationship.**

A similar extension methodology is used to create SentiWordNet, a publicly available lexicon specifically developed for opinion mining (sentiment analysis) purposes. **SentiWordNet** assigns to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity. More about SentiWordNet can be found at **sentiwordnet.isti.cnr.it.**

Another extension to WordNet is WordNet-Affect, developed by Strapparava and Valitutti (2004). They label WordNet synsets using affective labels representing different affective categories (emotion, cognitive state, attitude, and feeling). WordNet has also been directly used in sentiment analysis. For example, Kim and Hovy (2004) and Liu, Hu, and Cheng (2005) generate lexicons of positive and negative terms by starting with a small list of "seed" terms of known polarities (e.g., love, like, nice) and then using the antonymy and synonymy properties of terms to group them into either of the polarity categories.

Using a Collection of Training Documents

It is possible to perform sentiment classification using statistical analysis and machine-learning tools that take advantage of the vast resources of labeled (manually by annotators or using a star/point system) documents available. Product review Web sites such as Amazon, C-NET, eBay, RottenTomatoes, and the Internet Movie Database have all been extensively used as sources of annotated data. The star (or tomato, as it were) system provides an explicit label of the overall polarity of the review, and it is often taken as a gold standard in algorithm evaluation.

A variety of manually labeled textual data is available through evaluation efforts such as the Text REtrieval Conference, NII Test Collection for IR Systems, and Cross Language Evaluation Forum. The data sets these efforts produce often serve as a standard in the text mining community including sentiment analysis researchers. Individual researchers and research groups have also produced many interesting data sets. Technology Insights 7.2 lists some of the most popular ones. Once an already labeled textual data set has been obtained, a variety of predictive modeling and other machine-learning algorithms can be used to train sentiment classifiers. Some of the most popular algorithms used for this task include artificial neural networks, support vector machines, *k*-nearest neighbor, Naïve Bayes, decision trees, and expectation maximization-based clustering.

TECHNOLOGY INSIGHTS 7.2 Large Textual Data Sets for Predictive Text Mining and Sentiment Analysis

Following are a few of the most commonly used examples to large textual data sets:

Congressional Floor-Debate Transcripts: Published by Thomas, Pang, and Lee (2006); contains political speeches that are labeled to indicate whether the speaker supported or opposed the legislation discussed.

Economining: Published by the Stern School at New York University; consists of feedback postings for merchants at **Amazon.com**.

Cornell Movie-Review Data Sets: Introduced by Pang and Lee (2008); contains 1,000 positive and 1,000 negative automatically derived document-level labels and 5,331 positive and 5,331 negative sentences/snippets.

Stanford—Large Movie Review Data Set: A set of 25,000 highly polar movie reviews for training and 25,000 for testing. There is additional unlabeled data for use as well. Raw text and already processed bag-of-words formats are provided. (See **http://ai.stanford.edu/~amaas/data/sentiment**.)

MPQA Corpus: Corpus and Opinion Recognition System corpus; contains 535 manually annotated news articles from a variety of news sources containing labels for opinions and private states (beliefs, emotions, speculations, etc.).

Multiple-Aspect Restaurant Reviews: Introduced by Snyder and Barzilay (2007); contains 4,488 reviews with an explicit 1-to-5 rating for five different aspects: food, ambiance, service, value, and overall experience.

Identifying Semantic Orientation of Sentences and Phrases

Once the semantic orientation of individual words has been determined, it is often desirable to extend this to the phrase or sentence in which the word appears. The simplest way to accomplish such aggregation is to use some type of averaging for the polarities of words in the phrases or sentences. Though rarely applied, such aggregation can be as complex as using one or more machine-learning techniques to create a predictive relationship between the words (and their polarity values) and phrases or sentences.

Identifying Semantic Orientation of Documents

Even though the vast majority of the work in this area is done in determining semantic orientation of words and phrases/sentences, some tasks such as summarization and information retrieval could require semantic labeling of the whole document (Ramage et al., 2009). Similar to the case in aggregating sentiment polarity from word level to phrase or sentence level, aggregation to document level is also accomplished by some type of averaging. Sentiment orientation of the document might not make sense for very large documents; therefore, it is often used on small to medium-size documents posted on the Internet.

SECTION 7.6 REVIEW QUESTIONS

1. What is sentiment analysis? How does it relate to text mining?
2. What are the most popular application areas for sentiment analysis? Why?
3. What would be the expected benefits and beneficiaries of sentiment analysis in politics?
4. What are the main steps in carrying out sentiment analysis projects?
5. What are the two common methods for polarity identification? Explain.

7.7 WEB MINING OVERVIEW

The Internet has changed the landscape for conducting business forever. Because of the highly connected, flattened world and broadened competition field, today's companies are increasingly facing more opportunities (being able to reach customers and markets that they might never have thought possible) and more challenges (a globalized and ever-changing competitive marketplace). Companies with the vision and capabilities to deal with such a volatile environment are greatly benefiting from it, whereas others who resist adapting are having difficulty surviving. Having an engaged presence on the Internet is not a choice anymore; it is a business requirement. Customers are expecting companies to offer their products and/or services over the Internet. Customers are not only buying products and services but also talking about companies and sharing their transactional and usage experiences with others over the Internet.

The growth of the Internet and its enabling technologies has made data creation, data collection, and data/information/opinion exchange easier. Delays in service, manufacturing, shipping, delivery, and customer inquiries are no longer private incidents and are accepted as necessary evils. Now, thanks to social media tools and technologies on the Internet, everybody knows everything. Successful companies are the ones that embrace these Internet technologies and use them to improve their business processes to better communicate with their customers, understand their needs and wants, and serve them thoroughly and expeditiously. Being customer focused and keeping customers happy has never been as important a concept for businesses as they are now in this age of the Internet and social media.

The World Wide Web (or for short, Web) serves as an enormous repository of data and information on virtually everything one can conceive; business, personal, you name it—an abundant amount of it is there. The Web is perhaps the world's largest data and text repository, and the amount of information on the Web is growing rapidly. Much interesting information can be found online: whose home page is linked to which other pages, how many people have links to a specific Web page, and how a particular site is organized. In addition, each visitor to a Web site, each search on a search engine, each click on a link, and each transaction on an e-commerce site create additional data. Although unstructured textual data in the form of Web pages coded in HTML or XML are the dominant content of the Web, the Web infrastructure also contains hyperlink information (connections to other Web pages) and usage information (logs of visitors' interactions with Web sites), all of which provide rich data for knowledge discovery. Analysis of this information can help us make better use of Web sites and also aid us in enhancing relationships and value for the visitors to our own Web sites.

Because of its sheer size and complexity, mining the Web is not an easy undertaking by any means. The Web also poses great challenges for effective and efficient knowledge discovery (Han & Kamber, 2006):

- ***The Web is too big for effective data mining.*** The Web is so large and growing so rapidly that it is difficult to even quantify its size. Because of the sheer size of the Web, it is not feasible to set up a data warehouse to replicate, store, and integrate all of the data on the Web, making data collection and integration a challenge.
- ***The Web is too complex.*** The complexity of a Web page is far greater than that of a page in a traditional text document collection. Web pages lack a unified structure. They contain far more authoring style and content variation than any set of books, articles, or other traditional text-based document.
- ***The Web is too dynamic.*** The Web is a highly dynamic information source. Not only does the Web grow rapidly but also its content is constantly being updated. Blogs, news stories, stock market results, weather reports, sports scores, prices, company advertisements, and numerous other types of information are updated regularly on the Web.

- ***The Web is not specific to a domain.*** The Web serves a broad diversity of communities and connects billions of workstations. Web users have very different backgrounds, interests, and usage purposes. Most users might not have good knowledge of the structure of the information network and might not be aware of the heavy cost of a particular search that they perform.
- ***The Web has everything.*** Only a small portion of the information on the Web is truly relevant or useful to someone (or some task). It is said that 99 percent of the information on the Web is useless to 99 percent of Web users. Although this might not seem obvious, it is true that a particular person is generally interested in only a tiny portion of the Web, whereas the rest of the Web contains information that is uninteresting to the user and could swamp desired results. Finding the portion of the Web that is truly relevant to a person and the task being performed is a prominent issue in Web-related research.

These challenges have prompted many research efforts to enhance the effectiveness and efficiency of discovering and using data assets on the Web. A number of index-based Web search engines constantly search the Web and index Web pages under certain keywords. Using these search engines, an experienced user might be able to locate documents by providing a set of tightly constrained keywords or phrases. However, a simple keyword-based search engine suffers from several deficiencies. First, a topic of any breadth can easily contain hundreds or thousands of documents. This can lead to a large number of document entries returned by the search engine, many of which are marginally relevant to the topic. Second, many documents that are highly relevant to a topic might not contain the exact keywords defining them. As we cover in more detail later in this chapter, compared to keyword-based Web search, Web mining is a prominent (and more challenging) approach that can be used to substantially enhance the power of Web search engines because Web mining can identify authoritative Web pages, classify Web documents, and resolve many ambiguities and subtleties raised in keyword-based Web search engines.

Web mining (or Web data mining) is the process of discovering intrinsic relationships (i.e., interesting and useful information) from Web data, which are expressed in the form of textual, linkage, or usage information. The term *Web mining* was first used by Etzioni (1996); today, many conferences, journals, and books focus on Web data mining. It is a continually evolving area of technology and business practice. Web mining is essentially the same as data mining that uses data generated over the Web. The goal is to turn vast repositories of business transactions, customer interactions, and Web site usage data into actionable information (i.e., knowledge) to promote better decision making throughout the enterprise. Because of the increased popularity of the term *analytics*, today many have started to refer to Web mining as *Web analytics*. However, these two terms are not the same. Whereas Web analytics is primarily Web site usage focused data, Web mining is inclusive of all data generated via the Internet including transaction, social, and usage data. Where Web analytics aims to describe what has happened on the Web site (employing a predefined, metrics-driven descriptive analytics methodology), Web mining aims to discover previously unknown patterns and relationships (employing a novel predictive or prescriptive analytics methodology). From a big-picture perspective, Web analytics can be considered to be a part of Web mining. Figure 7.11 presents a simple taxonomy of Web mining divided into three main areas: Web content mining, Web structure mining, and Web usage mining. In the figure, the data sources used in these three main areas are also specified. Although these three areas are shown separately, as you will see in the following section, they are often used collectively and synergistically to address business problems and opportunities.

As Figure 7.11 indicates, Web mining relies heavily on data mining and text mining and their enabling tools and techniques, which we covered in detail early in this chapter

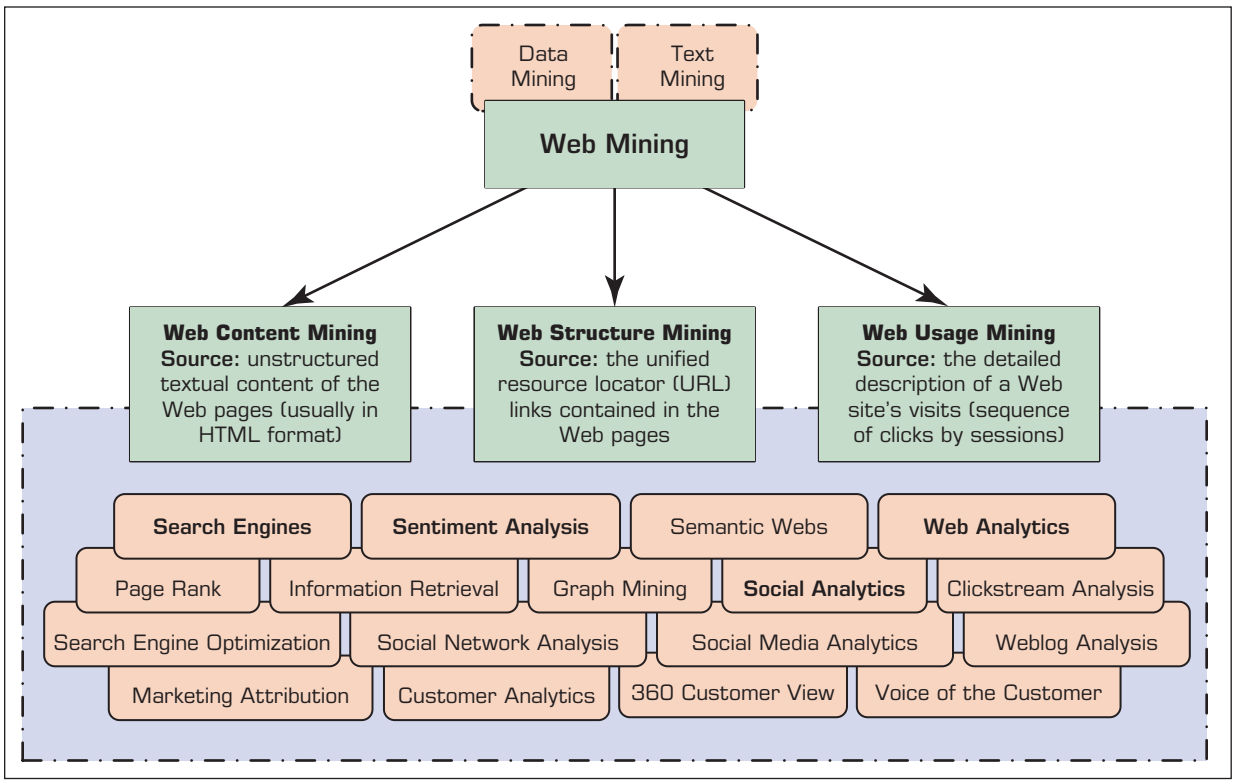


FIGURE 7.11 Simple Taxonomy of Web Mining.

and in the previous chapter (Chapter 4). The figure also indicates that these three generic areas are further extended into several very well-known application areas. Some of these areas were explained in previous chapters, and some of the others are covered in detail in this chapter.

Web Content and Web Structure Mining

Web content mining refers to the extraction of useful information from Web pages. The documents can be extracted in some machine-readable format so that automated techniques can extract some information from these Web pages. **Web crawlers** (also called **spiders**) are used to read through the content of a Web site automatically. The information gathered could include document characteristics similar to what is used in text mining, but it could also include additional concepts, such as the document hierarchy. Such an automated (or semiautomated) process of collecting and mining of Web content can be used for competitive intelligence (collecting intelligence about competitors' products, services, and customers). It can also be used for information/news/opinion collection and summarization, sentiment analysis, and automated data collection and structuring for predictive modeling. As an illustrative example of using Web content mining as an automated data collection tool, consider the following. For more than 10 years now, two of the three authors of this book (Drs. Sharda and Delen) have been developing models to predict the financial success of Hollywood movies before their theatrical release. The data that they use for training of the models come from several Web sites, each having a different hierarchical page structure. Collecting a large set of variables on thousands of movies (from the past several years) from these Web sites is a time-demanding,

error-prone process. Therefore, Sharda and Delen use Web content mining and spiders as an enabling technology to automatically collect, verify, validate (if the specific data item is available on more than one Web site, then the values are validated against each other and anomalies are captured and recorded), and store these values in a relational database. That way, they ensure the quality of the data while saving valuable time (days or weeks) in the process.

In addition to text, Web pages also contain hyperlinks pointing one page to another. Hyperlinks contain a significant amount of hidden human annotation that can potentially help to automatically infer the notion of centrality or *authority*. When a Web page developer includes a link pointing to another Web page, this could be regarded as the developer's endorsement of the other page. The collective endorsement of a given page by different developers on the Web might indicate the importance of the page and might naturally lead to the discovery of authoritative Web pages (Miller, 2005). Therefore, the vast amount of Web linkage information provides a rich collection of information about the relevance, quality, and structure of the Web's contents and thus is a rich source for Web mining.

Web content mining can also be used to enhance the results produced by search engines. In fact, search is perhaps the most prevailing application of Web content mining and Web structure mining. A search on the Web to obtain information on a specific topic (presented as a collection of keywords or a sentence) usually returns a few relevant, high-quality Web pages and a larger number of unusable Web pages. Use of a relevance index based on keywords and authoritative pages (or some measure of it) improves the search results and ranking of relevant pages. The idea of authority (or **authoritative pages**) stems from earlier information retrieval work using citations among journal articles to evaluate the impact of research papers (Miller, 2005). Although that was the origination of the idea, there are significant differences between the citations in research articles and hyperlinks on Web pages. First, not every hyperlink represents an endorsement (some links are created for navigation purposes and some are for paid advertisements). Although this is true, if the majority of the hyperlinks are of the endorsement type, then the collective opinion will still prevail. Second, for commercial and competitive interests, one authority rarely has its Web page point to rival authorities in the same domain. For example, Microsoft might prefer not to include links on its Web pages to Apple's Web sites because this could be regarded as an endorsement of its competitor's authority. Third, authoritative pages are seldom particularly descriptive. For example, the main Web page of Yahoo! might not contain the explicit self-description that it is in fact a Web search engine.

The structure of Web hyperlinks has led to another important category of Web pages called a **hub**, which is one or more Web pages that provide a collection of links to authoritative pages. Hub pages might not be prominent, and only a few links might point to them; however, hubs provide links to a collection of prominent sites on a specific topic of interest. A hub could be a list of recommended links on an individual's home page, recommended reference sites on a course Web page, or a professionally assembled resource list on a specific topic. Hub pages play the role of implicitly conferring the authorities on a narrow field. In essence, a close symbiotic relationship exists between good hubs and authoritative pages; a good hub is good because it points to many good authorities; and a good authority is good because it is being pointed to by many good hubs. Such relationships between hubs and authorities make it possible to automatically retrieve high-quality content from the Web.

The most popular publicly known and referenced algorithm used to calculate hubs and authorities is **hyperlink-induced topic search (HITS)**. It was originally developed by Kleinberg (1999) and has since been improved by many researchers. HITS is a link-analysis algorithm that rates Web pages using the hyperlink information contained within

them. In the context of Web search, the HITS algorithm collects a base document set for a specific query. It then recursively calculates the hub and authority values for each document. To gather the base document set, a root set that matches the query is fetched from a search engine. For each document retrieved, a set of documents that points to the original document and another set of documents that is pointed to by the original document are added to the set as the original document's neighborhood. A recursive process of document identification and link analysis continues until the hub and authority values converge. These values are then used to index and prioritize the document collection generated for a specific query.

Web structure mining is the process of extracting useful information from the links embedded in Web documents. It is used to identify authoritative pages and hubs, which are the cornerstones of the contemporary page-rank algorithms that are central to popular search engines such as Google and Yahoo! Just as links going to a Web page can indicate a site's popularity (or authority), links within the Web page (or the complete Web site) can indicate the depth of coverage of a specific topic. Analysis of links is very important in understanding the interrelationships among large numbers of Web pages, leading to a better understanding of a specific Web community, clan, or clique.

SECTION 7.7 REVIEW QUESTIONS

1. What are some of the main challenges the Web poses for knowledge discovery?
2. What is Web mining? How does it differ from regular data mining or text mining?
3. What are the three main areas of Web mining?
4. What is Web content mining? How can it be used for competitive advantage?
5. What is Web structure mining? How does it differ from Web content mining?

7.8 SEARCH ENGINES

In this day and age, there is no denying the importance of Internet search engines. As the size and complexity of the World Wide Web increase, finding what you want is becoming a complex and laborious process. People use search engines for a variety of reasons. We use them to learn about a product or service before committing to buy it (including who else is selling it, what the prices are at different locations/sellers, common issues people are discussing about it, how satisfied previous buyers are, and what other products or services might be better) and to search for places to go, people to meet, things to do. In a sense, search engines have become the centerpiece of most Internet-based transactions and other activities. The incredible success and popularity of Google, the most popular search engine company, is a good testament to this claim. What is somewhat of a mystery to many is how a search engine actually does what it is meant to do. In simplest terms, a **search engine** is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multiword terms, or a complete sentence) users have provided that have to do with the subject of their inquiry. Search engines are the workhorses of the Internet, responding to billions of queries in hundreds of different languages every day.

Technically speaking, "search engine" is the popular term for information retrieval systems. Although Web search engines are the most popular, search engines are often used in contexts other than the Web, such as desktop search engines and document search engines. As you will see in this section, many of the concepts and techniques that we covered in text analytics and text mining early in this chapter also apply here. The overall goal of a search engine is to return one or more documents/pages (if more than one document/page applies, then a ranked-order list is often provided) that best match the user's query. The two metrics that are often used to evaluate search engines are

effectiveness (or quality—finding the right documents/pages) and *efficiency* (or speed—returning a response quickly). These two metrics tend to work in reverse directions; improving one tends to worsen the other. Often, based on the user expectation, search engines focus on one at the expense of the other. Better search engines are the ones that excel in both at the same time. Because search engines not only search but also find and return documents/pages, perhaps a more appropriate name for them would have been *finding engines*.

Anatomy of a Search Engine

Now let us dissect a search engine and look inside it. At the highest level, a search engine system is composed of two main cycles: a development cycle and a responding cycle (see the structure of a typical Internet search engine in Figure 7.12). While one is interfacing with the World Wide Web, the other is interfacing with the user. One can think of the development cycle as a production process (manufacturing and inventorying documents/pages) and the responding cycle as a retailing process (providing customers/users what they want). In the following section, these two cycles are explained in more detail.

1. Development Cycle

The two main components of the development cycle are the Web crawler and document indexer. The purpose of this cycle is to create a huge database of documents/pages organized and indexed based on their content and information value. The reason for developing such a repository of documents/pages is quite obvious: Due to its sheer size and complexity, searching the Web to find pages in response to a user query is not practical (or feasible within a reasonable time frame); therefore, search engines "cache the Web" into their database and use the cached version of the Web for searching and finding. Once created, this database allows search engines to rapidly and accurately respond to user queries.

WEB CRAWLER A Web crawler (also called a *spider* or *Web spider*) is a piece of software that systematically browses (crawls through) the World Wide Web for the purpose of finding and fetching Web pages. Often Web crawlers copy all the pages they visit for later processing by other functions of a search engine.

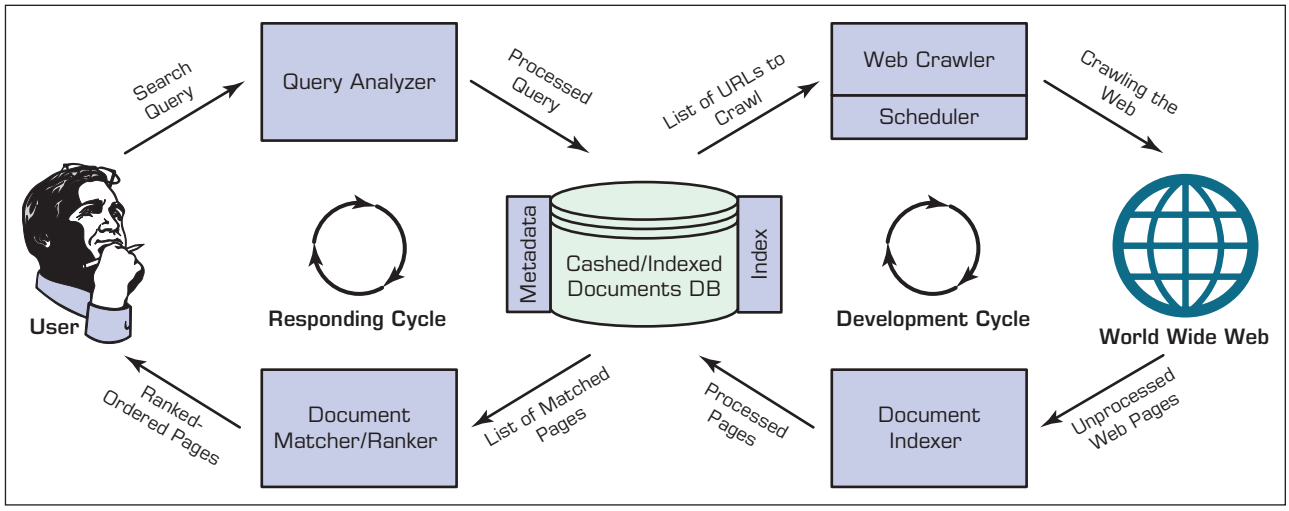


FIGURE 7.12 Structure of a Typical Internet Search Engine.

A Web crawler starts with a list of URLs to visit, which are listed in the scheduler and often are called the *seeds*. These URLs can come from submissions made by Webmasters or, more often, from the internal hyperlinks of previously crawled documents/pages. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit (i.e., the scheduler). URLs in the scheduler are recursively visited according to a set of policies determined by the specific search engine. Because there are large volumes of Web pages, the crawler can download only a limited number of them within a given time; therefore, it might need to prioritize its downloads.

DOCUMENT INDEXER As the documents are found and fetched by the crawler, they are stored in a temporary staging area for the document indexer to grab and process. The document indexer is responsible for processing the documents (Web pages or document files) and placing them into the document database. To convert the documents/pages into the desired, easily searchable format, the document indexer performs the following tasks.

STEP 1: PREPROCESSING THE DOCUMENTS Because the documents fetched by the crawler might all be in different formats, for the ease of processing them further, in this step they all are converted to some type of standard representation. For instance, different content types (text, hyperlink, image, etc.) could be separated from each other, formatted (if necessary), and stored in a place for further processing.

STEP 2: PARSING THE DOCUMENTS This step is essentially the application of text mining (i.e., computational linguistic, NLP) tools and techniques to a collection of documents/pages. In this step, first the standardized documents are parsed into components to identify index-worthy words/terms. Then, using a set of rules, the words/terms are indexed. More specifically, using tokenization rules, the words/terms/entities are extracted from the sentences in these documents. Using proper lexicons, the spelling errors and other anomalies in these words/terms are corrected. Not all the terms are discriminators. The nondiscriminating words/terms (also known as *stop words*) are eliminated from the list of index-worthy words/terms. Because the same word/term can be in many different forms, stemming is applied to reduce the words/terms to their root forms. Again, using lexicons and other language-specific resources (e.g., WordNet), synonyms and homonyms are identified, and the word/term collection is processed before moving into the indexing phase.

STEP 3: CREATING THE TERM-BY-DOCUMENT MATRIX In this step, the relationships between the words/terms and documents/pages are identified. The weight can be as simple as assigning 1 for presence or 0 for absence of the word/term in the document/page. Usually more sophisticated weight schemas are used. For instance, as opposed to binary, one can choose to assign frequency of occurrence (number of times the same word/term is found in a document) as a weight. As we saw earlier in this chapter, text mining research and practice have clearly indicated that the best weighting could come from the use of *term frequency* divided by inverse document frequency (TF/IDF). This algorithm measures the frequency of occurrence of each word/term within a document and then compares that frequency against the frequency of occurrence in the document collection. As we all know, not all high-frequency words/terms are good document discriminators, and a good document discriminator in a domain might not be one in another domain. Once the weighing schema is determined, the weights are calculated and the term-by-document index file is created.

2. Response Cycle

The two main components of the responding cycle are the query analyzer and the document matcher/ranker.

QUERY ANALYZER The query analyzer is responsible for receiving a search request from the user (via the search engine's Web server interface) and converting it into a standardized data structure so that it can be easily queried/matched against the entries in the document database. How the query analyzer does what it is supposed to do is quite similar to what the document indexer does (as we have just explained). The query analyzer parses the search string into individual words/terms using a series of tasks that include tokenization, removal of stop words, stemming, and word/term disambiguation (identification of spelling errors, synonyms, and homonyms). The close similarity between the query analyzer and the document indexer is not coincidental. In fact, it is quite logical because both are working off the document database; one is putting in documents/pages using a specific index structure, and the other is converting a query string into the same structure so that it can be used to quickly locate most relevant documents/pages.

DOCUMENT MATCHER/RANKER This is where the structured query data are matched against the document database to find the most relevant documents/pages and rank them in the order of relevance/importance. The proficiency of this step is perhaps the most important component when different search engines are compared to one another. Every search engine has its own (often proprietary) algorithm that it uses to carry out this important step.

The early search engines used a simple keyword match against the document database and returned a list of ordered documents/pages when the determinant of the order was a function that used the number of words/terms matched between the query and the document along with the weights of those words/terms. The quality and the usefulness of the search results were not all that good. Then, in 1997, the creators of Google came up with a new algorithm, called PageRank. As the name implies, PageRank is an algorithmic way to rank order documents/pages based on their relevance and value/importance. Even though PageRank is an innovative way to rank documents/pages, it is an augmentation to the process of retrieving relevant documents from the database and ranking them based on the weights of the words/terms. Google does all of these collectively and more to identify the most relevant list of documents/pages for a given search request. Once an ordered list of documents/pages is created, it is pushed back to the user in an easily digestible format. At this point, users might choose to click on any of the documents in the list, and it might not be the one at the top. If they click on a document/page link that is not at the top of the list, then can we assume that the search engine did not do a good job ranking them? Perhaps, yes. Leading search engines like Google monitor the performance of their search results by capturing, recording, and analyzing postdelivery user actions and experiences. These analyses often lead to more and more rules to further refine the ranking of the documents/pages so that the links at the top are more preferable to the end users.

Search Engine Optimization

Search engine optimization (SEO) is the intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results. In general, the higher it is ranked on the search results page, and the more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users. As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines, and which search engines are preferred by their targeted audience. Optimizing a Web site can involve editing its content, HTML, and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing

activities of search engines. Promoting a site to increase the number of backlinks, or inbound links, is another SEO tactic.

In the early days, in order to be indexed, the only thing that Webmasters needed to do was to submit the address of a page, or URL, to the various engines, which would then send a "spider" to "crawl" that page, extract links to other pages from it, and return information found on the page to the server for indexing. The process, as explained before, involves a search engine spider downloading a page and storing it on the search engine's own server, where a second program, known as an indexer, extracts various information about the page, such as the words it contains and where they are located as well as any weight for specific words, and all links the page contains, which are then placed into a scheduler for crawling at a later date. Today, search engines are no longer relying on Webmasters submitting URLs (even though they still can); instead, search engines are proactively and continuously crawling the Web and finding, fetching, and indexing everything about it.

Being indexed by search engines such as Google, Bing, and Yahoo! is not good enough for businesses. Getting ranked on the most widely used search engines (see Technology Insights 7.3 for a list of most widely used search engines) and getting ranked higher than your competitors are what make a difference in the eye of the customers and other constituents. A variety of methods can increase the ranking of a Web page within the search results. Cross-linking between pages of the same Web site to provide more links to the most important pages could improve its visibility. Writing content that includes frequently searched keyword phrases to be relevant to a wide variety of search queries tends to increase traffic. Updating content to keep search engines crawling back frequently can give additional weight to a site. Adding relevant keywords to a Web page's metadata, including the title tag and metadescription, will tend to improve the relevancy of a site's search listings, thus increasing traffic. URL normalization of Web pages (so that they are accessible via multiple and simpler URLs) and using canonical link elements and redirects can help make sure that the links to different versions of the Web pages and their URLs all count toward the Web site's link popularity score.

Methods for Search Engine Optimization

In general, SEO techniques can be classified into two broad categories: techniques that search engines recommend as part of good site design and those techniques of which search engines do not approve. The search engines attempt to minimize the effect of the latter, which is often called *spamdexing* (also known as *search spam, search engine spam*, or *search engine poisoning*). Industry commentators, and the practitioners who employ them, have classified these methods as either white-hat SEO or black-hat SEO (Goodman, 2005). White hats tend to produce results that last a long time, whereas black hats anticipate that their sites might eventually be banned either temporarily or permanently once the search engines discover what they are doing.

An SEO technique is considered white hat if it conforms to the search engine's guidelines and involves no deception. Because search engine guidelines are not written as a series of rules or commandments, this is an important distinction to note. White-hat SEO is not just about following guidelines but also about ensuring that the content a search engine indexes and subsequently ranks is the same content a user will see. White-hat advice is generally summed up as creating content for users, not for search engines, and then making that content easily accessible to the spiders rather than attempting to trick the algorithm from its intended purpose. White-hat SEO is in many ways similar to Web development that promotes accessibility, although the two are not identical.

TECHNOLOGY INSIGHTS 7.3 Top 15 Most Popular Search Engines (August 2016)

These are the 15 most popular search engines as derived from eBizMBA Rank (**ebizmba.com/articles/search-engines**), which is a constantly updated average of each Web site's *Alexa* Global Traffic Rank and U.S. Traffic Rank from both Compete and Quantcast.

Rank	Name	Estimated Unique Monthly Visitors
1	Google	1,600,000,000
2	Bing	400,000,000
3	Yahoo! Search	300,000,000
4	Ask	245,000,000
5	AOL Search	125,000,000
6	Wow	100,000,000
7	WebCrawler	65,000,000
8	MyWebSearch	60,000,000
9	Infospace	24,000,000
10	Info	13,500,000
11	DuckDuckGo	11,000,000
12	Contenko	10,500,000
13	Dogpile	7,500,000
14	Alhea	4,000,000
15	ixQuick	1,000,000

Black-hat SEO attempts to improve rankings in ways that are not approved by the search engines or involve deception. One black-hat technique uses text that is hidden, either as text colored similar to the background, in an invisible div tag (that defines a division or a section in an HTML document), or positioned off-screen. Another method gives a different page depending on whether the page is being requested by a human visitor or a search engine, a technique known as *cloaking*. Search engines can penalize sites they discover using black-hat methods, either by reducing their rankings or eliminating their listings from their databases altogether. Such penalties can be applied either automatically by the search engines' algorithms or by a manual site review. One example was the February 2006 Google removal of both BMW Germany and Ricoh Germany for use of unapproved practices (Cutts, 2006). Both companies, however, quickly apologized, fixed their practices, and were restored to Google's list.

For some businesses, SEO can generate a significant return on investment. However, one should keep in mind that search engines are not paid for organic search traffic, their algorithms change constantly, and there are no guarantees of continued referrals. Due to this lack of certainty and stability, a business that relies heavily on search engine traffic can suffer major losses if the search engine decides to change its algorithms and stop sending visitors. According to Google's CEO, Eric Schmidt, in 2010, Google made over 500 algorithm changes—almost 1.5 per day. Because of the difficulty in keeping up with changing search engine rules, companies that rely on search traffic practice one or more of the following: (1) hire a company that specializes in SEO (there seem to be an abundant number of those today) to continuously improve your site's appeal to changing practices of the search engines, (2) pay the search engine providers to be listed on the paid sponsors' sections, and (3) consider liberating yourself from dependence on search engine traffic.

Either originating from a search engine (organically or otherwise) or coming from other sites and places, what is most important for an e-commerce site is to maximize the likelihood of customer transactions. Having many visitors without sales is not what a typical e-commerce site is built for. Application Case 7.7 discusses a large Internet-based shopping mall where detailed analysis of customer behavior (using clickstreams and other data sources) is used to significantly improve the conversion rate.

Either originating from a search engine (organically or otherwise), responding to email-based marketing campaigns, or coming from social media sites, what is most important for an e-commerce site is to maximize its leads and subsequent customer sales transactions. Application Case 7.7 shows how a century-old fashionable cloth and accessory company used email-based campaigns to generate large number of new leads for its e-commerce business.

Application Case 7.7

Delivering Individualized Content and Driving Digital Engagement: How Barbour Collected More Than 49,000 New Leads in One Month with Teradata Interactive

Background

Founded in 1894, Barbour is an English heritage and lifestyle brand renowned for its waterproof outerwear—especially its classic waxed-cotton jacket. With more than 10,000 jackets ordered and hand-made each year, Barbour has held a strong position in the luxury goods industry for more than a century, building a strong relationship with fashion-conscious men and women of the British countryside. In 2000, Barbour broadened its product offering to include a full lifestyle range of everyday clothes and accessories. Its major markets are the United Kingdom, the United States, and Germany; however, Barbour holds a presence in more than 40 countries worldwide, including Austria, New Zealand, and Japan. Using individualized insights derived with the services and digital marketing capabilities of Teradata Interactive, Barbour ran a one-month campaign that generated 49,700 new leads and 450,000 clicks to its website.

The Challenge: Taking Ownership of Customer Relationships

Barbour has experienced outstanding consistent growth within its lifetime, and in August 2013, it launched its first e-commerce site in a bid to gain a stronger online presence. However, being a late starter in the e-commerce world, it was a challenge for Barbour to establish itself in the saturated digital arena. Having previously sold its products only through wholesalers and independent retail resellers, Barbour wanted to take ownership of the end-user relationship, whole customer journey, and perception of the brand. While the brand is iconic and highly respected around the world, Barbour was aware of the importance in establishing direct relationships with its target audience—especially when encouraging users to engage with its new e-commerce platform. It also understood that it needed to take more control of shaping the customer journey. That way Barbour could create and maintain the same exceptional level of quality in the user experience as that applied to the manufacturing of its products. To do this, the company needed to develop its understanding of its target market's online behavior. With the goal of reaching its target audience in order to build meaningful customer relationships, Barbour approached Teradata. Barbour's marketing department needed Teradata Interactive to offer a solution that would increase its knowledge of the unique characteristics and needs of its individual customers, as well as support the launch of its new UK e-commerce website.

The Solution: Implementing a Lead Nurture Program

The increasing shift to global e-commerce and the growth in digital consumerism require brands to hold a strong online presence. This also means that retailers have to implement strategies that support their customers' evolving wants and needs, online and offline. Barbour and Teradata Interactive embarked on the design and construction of a Lead Nurture Program that ran over a one-month period. The campaign objective was to not only raise awareness and create demand for immediate sales activity but also to create a more long-term engagement mechanism that would lead to more

(*Continued*)

Application Case 7.7 (Continued)

sales over a sustained period of time. It was clear from the start that the strong relationship Barbour enjoys with its customers was a crucial factor that set it apart from its luxury retail competitors. Teradata Interactive was keen to ensure this relationship was respected through the lead generation process.

The execution of the campaign was unique to Barbour. Typical lead generation campaigns were often executed as single registration events with a single sales promotion in mind. The data were usually restricted to just email addresses and basic profile fields, generated without consideration of the registrant's personal needs and imported to be used solely for generic newsletter campaigns. This strategy often missed a huge opportunity for brands when learning about their prospects, often resulting in poor sales conversions. Teradata Interactive understood that the true value of lead generation is twofold. First of all, by using the registration event to gather as much information as possible, the understanding of future buying intent and its affecting factors are developed. Second, by making sure that the collated data are effectively used to deliver valuable and individualized content, relevant sales opportunities are provided to the customer when they are next in the market to buy. To make sure this strategy drove long-term sales, Teradata Interactive built a customer lifecycle program which delivered content over email and online display.

The nurture program content was integrated with display advertising and encouraged social media sharing. With Teradata Interactive's smart tagging of nurture content, Barbour was able to segment audiences according to their product preferences and launch display re-targeting banners. Registrants were also invited to share content socially, which enabled Teradata Interactive to identify "social propensity" and segment users for future loyalty schemes and "Tell-a-Friend" activities. In addition to the focus of increasing Barbour's newsletter base, Teradata conducted a data audit to analyze all of the data collected and better understand what factors would influence user engagement behavior.

Results

The strong collaboration between Teradata and Barbour meant that over the one-month campaign period, Barbour was able to create new and innovative ways of communicating with its customers. More than 49,700 leads were collected within the UK and DACH regions, and the lead generation program showed open rates of up to 60% and click-through-rates of between 4 and 11 percent. The campaign also generated 450,000+ clicks to Barbour's website and was so popular with fashion bloggers and national press that it was featured as a story in *The Daily Mirror*. Though the campaign was only a month long, a key focus was to help Barbour's future marketing strategy. A preference center survey was implemented into the campaign design, which resulted in a 65 percent incentivized completion rate. User data included:

- Social network engagement
- Device engagement
- Location to nearest store
- Important considerations to the customer

A deep level of insight has effectively given Barbour a huge capability to deliver personalized content and offers to its user base.

Questions for Case 7.7

1. What does Barbour do? What was the challenge Barbour was facing?
2. What was the proposed analytics solution?
3. What were the results?

Source: Teradata Case Study, "How Barbour Collected More Than 49,000 New Leads in One Month with Teradata Interactive" **http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB-8791_Interactive-Case-Study_Barbour.pdf** (accessed November 2018).

SECTION 7.8 REVIEW QUESTIONS

1. What is a search engine? Why are search engines critically important for today's businesses?
2. What is a Web crawler? What is it used for? How does it work?
3. What is "search engine optimization"? Who benefits from it?
4. What things can help Web pages rank higher in search engine results?

7.9 WEB USAGE MINING (WEB ANALYTICS)

Web usage mining (also called **Web analytics**) is the extraction of useful information from data generated through Web page visits and transactions. Analysis of the information collected by Web servers can help us better understand user behavior. Analysis of these data is often called **clickstream analysis**. By using data and text mining techniques, a company might be able to discern interesting patterns from the clickstreams. For example, it might learn that 60 percent of visitors who searched for "hotels in Maui" had searched earlier for "airfares to Maui." Such information could be useful in determining where to place online advertisements. Clickstream analysis might also be useful for knowing *when* visitors access a site. For example, if a company knew that 70 percent of software downloads from its Web site occurred between 7 and 11 P.M., it could plan for better customer support and network bandwidth during those hours. Figure 7.13 shows the process of extracting knowledge from clickstream data and how the generated knowledge is used to improve the process, improve the Web site, and most importantly, increase the customer value.

Web Analytics Technologies

There are numerous tools and technologies for Web analytics in the marketplace. Because of their power to measure, collect, and analyze Internet data to better understand and optimize Web usage, the popularity of Web analytics tools is increasing. Web analytics holds the promise of revolutionizing how business is done on the Web. Web analytics is not just a tool for measuring Web traffic; it can also be used as a tool for e-business and market research and to assess and improve the effectiveness of e-commerce Web sites. Web analytics applications can also help companies measure the results of traditional print or broadcast advertising campaigns. It can help estimate how traffic to a Web site changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a Web site and the number of page views. It helps gauge traffic and popularity trends, which can be used for market research.

There are two main categories of Web analytics: off-site and on-site. Off-site Web analytics refers to Web measurement and analysis about you and your products that take place outside your Web site. These measurements include a Web site's potential audience (prospect or opportunity), share of voice (visibility or word of mouth), and buzz (comments or opinions) that is happening on the Internet.

What is more mainstream has been on-site Web analytics. Historically, Web analytics has been referred to as on-site visitor measurement. However, in recent years, this has blurred, mainly because vendors are producing tools that span both categories. On-site Web

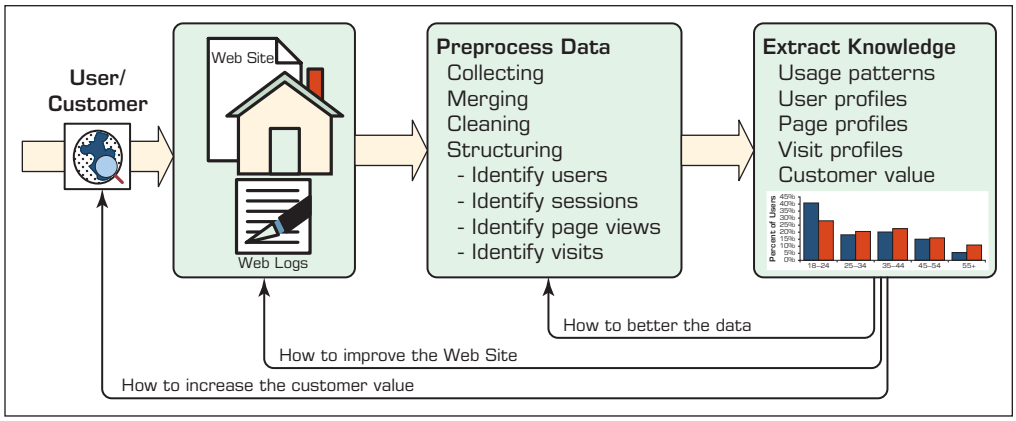


FIGURE 7.13 Extraction of Knowledge from Web Usage Data.

analytics measure visitors' behavior once they are on your Web site. This includes its drivers and conversions—for example, the degree to which different landing pages are associated with online purchases. On-site Web analytics measure the performance of your Web site in a commercial context. The data collected on the Web site is then compared against key performance indicators for performance and used to improve audience response on a Web site or marketing campaign. Even though Google Analytics is the most widely used on-site Web analytics service, others are provided by Yahoo! and Microsoft, and newer and better tools are emerging constantly that provide additional layers of information.

There are two technical ways to collect the data with on-site Web analytics. The first and more traditional method is the server log file analysis by which the Web server records file requests made by browsers. The second method is page tagging, which uses JavaScript embedded in the site page code to make image requests to a third-party analytics–dedicated server whenever a page is rendered by a Web browser (or when a mouse click occurs). Both collect data that can be processed to produce Web traffic reports. In addition to these two main streams, other data sources can also be added to augment Web site behavior data. These other sources can include e-mail, direct mail campaign data, sales and lead history, or social media–originated data.

Web Analytics Metrics

Using a variety of data sources, Web analytics programs provide access to much valuable marketing data, which can be leveraged for better insights to grow your business and better document your return on investment (ROI). The insight and intelligence gained from Web analytics can be used to effectively manage the marketing efforts of an organization and its various products or services. Web analytics programs provide nearly real-time data, which can document an organization's marketing campaign successes or empower it to make timely adjustments to its current marketing strategies.

Whereas Web analytics provides a broad range of metrics, four categories of metrics are generally actionable and can directly impact your business objectives (The Westover Group, 2013). These categories include:

- Web site usability: How were they using my Web site?
- Traffic sources: Where did they come from?
- Visitor profiles: What do my visitors look like?
- Conversion statistics: What does it all mean for the business?

Web Site Usability

Beginning with your Web site, let's take a look at how well it works for your visitors. This is where you can learn how "user friendly" it really is or whether or not you are providing the right content.

1. ***Page views.*** The most basic of measurements, this one is usually presented as the "average page views per visitor." If people come to your Web site and do not view many pages, then your Web site could have issues with its design or structure. Another explanation for low page views is a disconnect in the marketing messages that brought the visitor to the site and the content that is actually available.
2. ***Time on site.*** Similar to page views, this is a fundamental measurement of a visitor's interaction with your Web site. Generally, the longer a person spends on your Web site, the better it is. That could mean they are carefully reviewing your content, utilizing interactive components you have available, and building toward an informed decision to buy, respond, or take the next step you have provided. On the contrary, the time on site also needs to be examined against the number of pages viewed to make sure the visitor is not spending his or her time trying to locate content that should be more readily accessible.

3. ***Downloads.*** This includes PDFs, videos, and other resources you make available to your visitors. Consider how accessible these items are as well as how well they are promoted. If your Web statistics, for example, reveal that 60 percent of the individuals who watch a demo video also make a purchase, then you will want to strategize to increase viewership of that video.
4. ***Click map.*** Most analytics programs can show you the percentage of clicks each item on your Web page received. This includes clickable photos, text links in your copy, downloads, and, of course, any navigation you have on the page. Are they clicking the most important items?
5. ***Click paths.*** Although an assessment of click paths is more involved, they can quickly reveal where you might be losing visitors in a specific process. A well-designed Web site uses a combination of graphics and information architecture to encourage visitors to follow "predefined" paths through your Web site. These are not rigid pathways but intuitive steps that align with the various processes you have built into the Web site. One process might be that of "educating" a visitor who has minimum understanding of your product or service. Another might be a process of "motivating" a returning visitor to consider an upgrade or repurchase. A third process might be structured around items you market online. You will have as many process pathways in your Web site as you have target audiences, products, and services. Each can be measured through Web analytics to determine how effective it is.

Traffic Sources

Your Web analytics program is an incredible tool for identifying where your Web traffic originates. Basic categories such as search engines, referral Web sites, and visits from bookmarked pages (i.e., direct) are compiled with little involvement by the marketer. With a small amount of effort, however, you can also identify Web traffic that was generated by your various offline or online advertising campaigns.

1. ***Referral Web sites.*** Other Web sites that contain links that send visitors directly to your Web site are considered referral Web sites. Your analytics program will identify each referral site your traffic comes from, and a deeper analysis will help you determine which referrals produce the greatest volume, the highest conversions, the most new visitors, and so on.
2. ***Search engines.*** Data in the search engine category is divided between paid search and organic (or natural) search. You can review the top keywords that generated Web traffic to your site and see if they are representative of your products and services. Depending upon your business, you might want to have hundreds (or thousands) of keywords that draw potential customers. Even the simplest product search can have multiple variations based on how the individual phrases the search query.
3. ***Direct.*** Direct searches are attributed to two sources. Individuals who bookmark one of your Web pages in their favorites and click that link will be recorded as a direct search. Another source occurs when someone types your URL directly into her or his browser. This happens when someone retrieves your URL from a business card, brochure, print ad, radio commercial, and so on. That's why it is a good strategy to use coded URLs.
4. ***Offline campaigns.*** If you utilize advertising options other than Web-based campaigns, your Web analytics program can capture performance data if you include a mechanism for sending them to your Web site. Typically, this is a dedicated URL that you include in your advertisement (i.e., "**www.mycompany.com/offer50**") that delivers those visitors to a specific landing page. You now have data on how many responded to that ad by visiting your Web site.

5. ***Online campaigns.*** If you are running a banner ad campaign, search engine advertising campaign, or even e-mail campaign, you can measure individual campaign effectiveness by simply using a dedicated URL similar to the offline campaign strategy.

Visitor Profiles

One of the ways you can leverage your Web analytics into a really powerful marketing tool is through segmentation. By blending data from different analytics reports, you will begin to see a variety of user profiles emerge.

1. ***Keywords.*** Within your analytics report, you can see what keywords visitors used in search engines to locate your Web site. If you aggregate your keywords by similar attributes, you will begin to see distinct visitor groups who are using your Web site. For example, the particular search phrase that was used can indicate how well they understand your product or its benefits. If they use words that mirror your own product or service descriptions, then they probably are already aware of your offerings from effective advertisements, brochures, and so on. If the terms are more general in nature, then your visitors are seeking a solution for a problem that has happened upon your Web site. If this second group of searchers is sizable, then you will want to ensure that your site has a strong education component to convince them they have found their answer and then move them into your sales channel.
2. ***Content groupings.*** Depending on how you group your content, you could be able to analyze sections of your Web site that correspond to specific products, services, campaigns, and other marketing tactics. If you conduct a number of trade shows and drive traffic to your Web site for specific product literature, then your Web analytics will highlight the activity in that section.
3. ***Geography.*** Analytics permits you to see where your traffic geographically originates, including country, state, and city locations. This can be especially useful if you use geotargeted campaigns or want to measure your visibility across a region.
4. ***Time of day.*** Web traffic generally has peaks at the beginning of the workday, during lunch, and toward the end of the workday. It is not unusual, however, to find strong Web traffic entering your Web site up until the late evening. You can analyze these data to determine when people browse versus buy and to make decisions on what hours you should offer customer service.
5. ***Landing page profiles.*** If you structure your various advertising campaigns properly, you can drive each of your targeted groups to a different landing page, which your Web analytics will capture and measure. By combining these numbers with the demographics of your campaign media, you can know what percentage of your visitors fits each demographic.

Conversion Statistics

Each organization defines a "conversion" according to its specific marketing objectives. Some Web analytics programs use the term *goal* to benchmark certain Web site objectives, whether that be a certain number of visitors to a page, a completed registration form, or an online purchase.

1. ***New visitors.*** If you are working to increase visibility, you will want to study the trends in your new visitors data. Analytics identifies all visitors as either new or returning.
2. ***Returning visitors.*** If you are involved in loyalty programs or offer a product that has a long purchase cycle, then your returning visitors data will help you measure progress in this area.

3. ***Leads.*** Once a form is submitted and a thank-you page is generated, you have created a lead. Web analytics will permit you to calculate a completion rate (or abandonment rate) by dividing the number of completed forms by the number of Web visitors that came to your page. A low completion percentage would indicate a page that needs attention.
4. ***Sales/conversions.*** Depending on the intent of your Web site, you can define a "sale" by an online purchase, a completed registration, an online submission, or any number of other Web activities. Monitoring these figures will alert you to any changes (or successes!) that occur further upstream.
5. ***Abandonment/exit rates.*** Just as important as those moving through your Web site are those who began a process and quit or came to your Web site and left after a page or two. In the first case, you'll want to analyze where the visitor terminated the process and whether there are a number of visitors quitting at the same place. Then investigate the situation for resolution. In the latter case, a high exit rate on a Web site or a specific page generally indicates an issue with expectations. Visitors click to your Web site based on some message contained in an advertisement, a presentation, and so on, and expect some continuity in that message. Make sure you are advertising a message that your Web site can reinforce and deliver.

Within each of these items are metrics that can be established for your specific organization. You can create a weekly dashboard that includes specific numbers or percentages that will indicate where you are succeeding—or highlight a marketing challenge that should be addressed. When these metrics are evaluated consistently and used in conjunction with other available marketing data, they can lead you to a highly quantified marketing program. Figure 7.14 shows a Web analytics dashboard created with freely available Google Analytics tools.

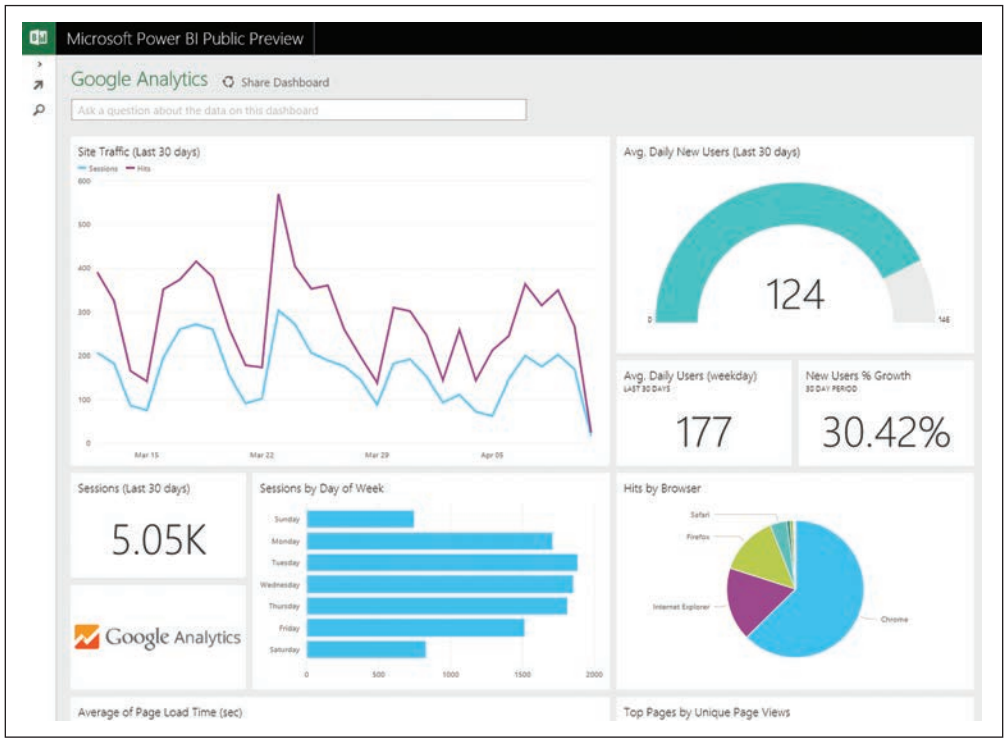


FIGURE 7.14 **Sample Web Analytics Dashboard.**

SECTION 7.9 REVIEW QUESTIONS

1. What are the three types of data generated through Web page visits?
2. What is clickstream analysis? What is it used for?
3. What are the main applications of Web mining?
4. What are commonly used Web analytics metrics? What is the importance of metrics?

7.10 SOCIAL ANALYTICS

Social analytics could mean different things to different people based on their worldview and field of study. For instance, the dictionary definition of *social analytics* refers to a philosophical perspective developed by the Danish historian and philosopher Lars-Henrik Schmidt in the 1980s. The theoretical object of the perspective is *socius*, a kind of "commonness" that is neither a universal account nor a communality shared by every member of a body (Schmidt, 1996). Thus, social analytics differs from traditional philosophy as well as sociology. It might be viewed as a perspective that attempts to articulate the contentions between philosophy and sociology.

Our definition of social analytics is somewhat different; as opposed to focusing on the "social" part (as is done in its philosophical definition), we are more interested in the "analytics" part of the term. Gartner (a very well-known global IT consultancy company) defined social analytics as "monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content" (**gartner.com/it-glossary/social-analytics/**). Social analytics include mining the textual content created in social media (e.g., sentiment analysis, NLP) and analyzing socially established networks (e.g., influencer identification, profiling, prediction) for the purpose of gaining insight about existing and potential customers' current and future behaviors, and about the likes and dislikes toward a firm's products and services. Based on this definition and the current practices, social analytics can be classified into two different, but not necessarily mutually exclusive, branches: social network analysis (SNA) and social media analytics.

Social Network Analysis

A **social network** is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships. The social network perspective provides a holistic approach to analyzing the structure and dynamics of social entities. The study of these structures uses SNA to identify local and global patterns, locate influential entities, and examine network dynamics. Social networks and their analysis is essentially an interdisciplinary field that emerged from social psychology, sociology, statistics, and graph theory. Development and formalization of the mathematical extent of SNA dates back to the 1950s; the development of foundational theories and methods of social networks dates back to the 1980s (Scott & Davis, 2003). SNA is now one of the major paradigms in business analytics, consumer intelligence, and contemporary sociology and is employed in a number of other social and formal sciences.

A social network is a theoretical construct useful in the social sciences to study relationships between individuals, groups, organizations, or even entire societies (social units). The term is used to describe a social structure determined by such interactions. The ties through which any given social unit connects represent the convergence of the various social contacts of that unit. In general, social networks are self-organizing, emergent, and complex, such that a globally coherent pattern appears from the local interaction of the elements (individuals and groups of individuals) that make up the system.

Following are a few typical social network types that are relevant to business activities.

COMMUNICATION NETWORKS Communication studies are often considered a part of both the social sciences and the humanities, drawing heavily on fields such as sociology, psychology, anthropology, information science, biology, political science, and economics. Many communications concepts describe the transfer of information from one source to another and thus can be represented as a social network. Telecommunication companies are tapping into this rich information source to optimize their business practices and to improve customer relationships.

COMMUNITY NETWORKS Traditionally, *community* has referred to a specific geographic location, and studies of community ties had to do with who talked, associated, traded, and attended social activities with whom. Today, however, there are extended "online" communities developed through social networking tools and telecommunications devices. Such tools and devices continuously generate large amounts of data that companies can use to discover invaluable, actionable information.

CRIMINAL NETWORKS In criminology and urban sociology, much attention has been paid to the social networks among criminal actors. For example, studying gang murders and other illegal activities as a series of exchanges between gangs can lead to better understanding and prevention of such criminal activities. Now that we live in a highly connected world (thanks to the Internet), much of the criminal networks' formations and their activities are being watched/pursued by security agencies using state-of-the-art Internet tools and tactics. Even though the Internet has changed the landscape for criminal networks and law enforcement agencies, the traditional social and philosophical theories still apply to a large extent.

INNOVATION NETWORKS Business studies on the diffusion of ideas and innovations in a network environment focus on the spread and use of ideas among the members of the social network. The idea is to understand why some networks are more innovative, and why some communities are early adopters of ideas and innovations (i.e., examining the impact of social network structure on influencing the spread of an innovation and innovative behavior).

Social Network Analysis Metrics

SNA, the systematic examination of social networks, views social relationships in terms of network theory consisting of nodes (representing individuals or organizations within the network) and ties/connections (which represent relationships between the individuals or organizations, such as friendship, kinship, or organizational position). These networks are often represented using social network diagrams, where nodes are represented as points and ties are represented as lines.

Application Case 7.8 provides an interesting example of multichannel social analytics.

Application Case 7.8

Tito's Vodka Establishes Brand Loyalty with an Authentic Social Strategy

If Tito's Handmade Vodka had to identify a single social media metric that most accurately reflects its mission, it would be engagement. Connecting with vodka lovers in an inclusive, authentic way is something Tito's takes very seriously, and the brand's social strategy reflects that vision.

Founded nearly two decades ago, Tito's credits the advent of social media with playing an integral role in engaging fans and raising brand awareness. In an interview with *Entrepreneur*, founder Bert "Tito" Beveridge credited social media for enabling Tito's to compete for shelf space with more established liquor

(*Continued*)

Application Case 7.8 (Continued)

brands. "Social media is a great platform for a word-of-mouth brand, because it's not just about who has the biggest megaphone," Beveridge told *Entrepreneur*.

As Tito's has matured, the social team has remained true to the brand's founding values and actively uses Twitter and Instagram to have one-on-one conversations and connect with brand enthusiasts. "We never viewed social media as another way to advertise," said Katy Gelhausen, Web & social media coordinator. "We're on social so our customers can talk to us."

To that end, Tito's uses Sprout Social to understand the industry atmosphere, develop a consistent social brand, and create a dialogue with its audience. As a result, Tito's recently organically grew its Twitter and Instagram communities by 43.5 percent and 12.6 percent, respectively, within four months.

Informing a Seasonal, Integrated Marketing Strategy

Tito's quarterly cocktail program is a key part of the brand's integrated marketing strategy. Each quarter, a cocktail recipe is developed and distributed through Tito's online and offline marketing initiatives.

It is important for Tito's to ensure that the recipe is aligned with the brand's focus as well as the larger industry direction. Therefore, Gelhausen uses Sprout's Brand Keywords to monitor industry trends and cocktail flavor profiles. "Sprout has been a really important tool for social monitoring. The Inbox is a nice way to keep on top of hashtags and see general trends in one stream," she said.

The information learned is presented to Tito's in-house mixology team and used to ensure that the same quarterly recipe is communicated to the brand's sales team and across marketing channels. "Whether you're drinking Tito's at a bar, buying it from a liquor store or following us on social media you're getting the same quarterly cocktail," said Gelhausen.

The program ensures that, at every consumer touch point, a person receives a consistent brand experience—and that consistency is vital. In fact, according to an Infosys study on the omnichannel shopping experience, 34 percent of consumers attribute cross-channel consistency as a reason they spend more on a brand. Meanwhile, 39 percent cite inconsistency as reason enough to spend less.

At Tito's, gathering industry insights starts with social monitoring on Twitter and Instagram through Sprout. But the brand's social strategy does not stop there. Staying true to its roots, Tito's uses the platform on a daily basis to authentically connect with customers.

Used with permission of Sprout Social, Inc.

Sprout's Smart Inbox displays Tito's Twitter and Instagram accounts in a single, cohesive feed. This helps Gelhausen manage inbound messages and quickly identify which require a response.

"Sprout allows us to stay on top of the conversations we're having with our followers. I love how you can easily interact with content from multiple accounts in one place," she said.

Spreading the Word on Twitter

Tito's approach to Twitter is simple: engage in personal, one-on-one conversations with fans. Dialogue is a driving force for the brand, and over the course of four months, 88 percent of Tweets sent were replies to inbound messages.

Using Twitter as an open line of communication between Tito's and its fans resulted in a 162.2 percent increase in engagement and a 43.5 percent gain in followers. Even more impressively, Tito's ended the quarter with 538,306 organic impressions—an 81 percent rise. A similar strategy is applied to Instagram, which Tito's uses to strengthen and foster a relationship with fans by publishing photos and videos of new recipe ideas, brand events, and initiatives.

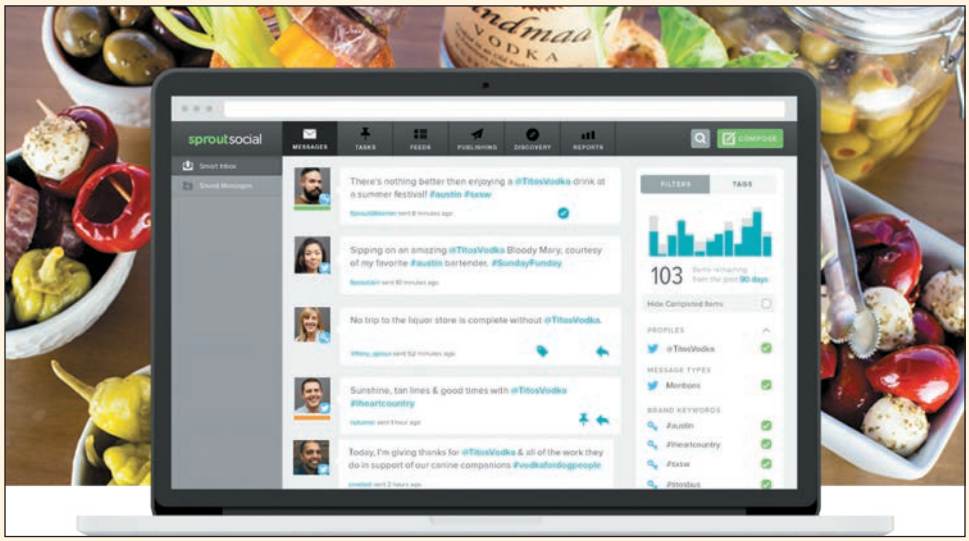
Capturing the Party on Instagram

On Instagram, Tito's primarily publishes lifestyle content and encourages followers to incorporate its brand in everyday occasions. Tito's also uses the platform to promote its cause through marketing efforts and to tell its brand story. The team finds value in Sprout's Instagram Profiles Report, which helps them identify what media is receiving the most engagement, analyze audience demographics and growth, dive more deeply into publishing patterns, and quantify outbound hashtag performance. "Given Instagram's new personalized feed, it's important that we pay attention to what really does resonate," said Gelhausen.

Using the Instagram Profiles Report, Tito's has been able to measure the impact of its Instagram marketing strategy and revise its approach accordingly. By utilizing the network as another way to engage with fans, the brand has steadily grown its organic audience. In four months, @TitosVodka saw a 12.6 percent rise in followers and a 37.1 percent increase in engagement. On average, each piece of published content gained 534 interactions, and mentions of the brand's hashtag, #titoshandmadevodka, grew by 33 percent.

Where to from Here?

Social is an ongoing investment in time and attention. Tito's will continue the momentum the brand experienced by segmenting each quarter into its own campaign. "We're always getting smarter with our social strategies and making sure that what we're posting is relevant and resonates," said Gelhausen. Using social to connect with fans in a consistent, genuine, and memorable way will remain a cornerstone of the brand's digital marketing efforts.



(*Continued*)

Application Case 7.8 (Continued)

Using Sprout's suite of social media management tools, Tito's will continue to foster a community of loyalists.

Some highlights of Tito's success follow:

- A 162 percent increase in organic engagement on Twitter.
- An 81 percent increase in organic Twitter impressions.
- A 37 percent increase in engagement on Instagram.

Questions for Case 7.8

1. How can social media analytics be used in the consumer products industry?
2. What do you think are the key challenges, potential solutions, and probable results in applying social media analytics in consumer products and services firms?

Source: SproutSocial Case Study, "Tito's Vodka Establishes Brand Loyalty with an Authentic Social Strategy." **http://sproutsocial.com/insights/case-studies/titos/** (accessed July 2016). Used with permission.

Over the years, various metrics (or measurements) have been developed to analyze social network structures from different perspectives. These metrics are often grouped into three categories: connections, distributions, and segmentation.

Connections

The connections category of metrics groups includes the following:

Homophily: The extent to which actors form ties with similar versus dissimilar others. Similarity can be defined by gender, race, age, occupation, educational achievement, status, values, or any other salient characteristic.

Multiplexity: The number of content forms contained in a tie. For example, two people who are friends and also work together would have a multiplexity of two. Multiplexity has been associated with relationship strength.

Mutuality/reciprocity: The extent to which two actors reciprocate each other's friendship or other interaction.

Network closure: A measure of the completeness of relational triads. An individual's assumption of network closure (i.e., that their friends are also friends) is called *transitivity*. Transitivity is an outcome of the individual or situational trait of need for cognitive closure.

Propinquity: The tendency for actors to have more ties with geographically close others.

Distributions

The following relate to the distributions category:

Bridge: An individual whose weak ties fill a structural hole, providing the only link between two individuals or clusters. It also includes the shortest route when a longer one is unfeasible due to a high risk of message distortion or delivery failure.

Centrality: A group of metrics that aims to quantify the importance or influence (in a variety of senses) of a particular node (or group) within a network. Examples of common methods of measuring centrality include betweenness centrality, closeness centrality, eigenvector centrality, alpha centrality, and degree centrality.

Density: The proportion of direct ties in a network relative to the total number possible.

Distance: The minimum number of ties required to connect two particular actors.

Structural holes: The absence of ties between two parts of a network. Finding and exploiting a structural hole can give an entrepreneur a competitive advantage. This concept was developed by sociologist Ronald Burt and is sometimes referred to as an *alternate conception of social capital*.

Tie strength: Defined by the linear combination of time, emotional intensity, intimacy, and reciprocity (i.e., mutuality). Strong ties are associated with homophily, propinquity, and transitivity, whereas weak ties are associated with bridges.

Segmentation

This category includes following:

Cliques and social circles: Groups are identified as *cliques* if every individual is directly tied to every other individual or *social circles* if there is less stringency of direct contact, which is imprecise, or as structurally cohesive blocks if precision is wanted.

Clustering coefficient: A measure of the likelihood that two members of a node are associates. A higher clustering coefficient indicates a greater *cliquishness*.

Cohesion: The degree to which actors are connected directly to each other by cohesive bonds. Structural cohesion refers to the minimum number of members who, if removed from a group, would disconnect the group.

Social Media Analytics

Social media refers to the enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks. Social media is a group of Internet-based software applications that build on the ideological and technological foundations of Web 2.0 and that allows the creation and exchange of user-generated content (Kaplan & Haenlein, 2010). Social media depends on mobile and other Web-based technologies to create highly interactive platforms for individuals and communities to share, co-create, discuss, and modify user-generated content. It introduces substantial changes to communication among organizations, communities, and individuals.

Since their emergence in the early 1990s, Web-based social media technologies have seen a significant improvement in both quality and quantity. These technologies take on many different forms, including online magazines, Internet forums, Web logs, social blogs, microblogging, wikis, social networks, podcasts, pictures, videos, and product/service evaluations/ratings. By applying a set of theories in the field of media research (social presence, media richness) and social processes (self-presentation, self-disclosure), Kaplan and Haenlein (2010) created a classification scheme with six different types of social media: collaborative projects (e.g., Wikipedia), blogs and microblogs (e.g., Twitter), content communities (e.g., YouTube), social networking sites (e.g., Facebook), virtual game worlds (e.g., World of Warcraft), and virtual social worlds (e.g., Second Life).

Web-based social media are different from traditional/industrial media, such as newspapers, television, and film, because they are comparatively inexpensive and accessible to enable anyone (even private individuals) to publish or access/consume information. Industrial media generally require significant resources to publish information because in most cases, the articles (or books) go through many revisions before being published (as was the case in the publication of this very book). The following are some of the most prevailing characteristics that help differentiate between social and industrial media (Morgan, Jones, & Hodges, 2010):

Quality: In industrial publishing—mediated by a publisher—the typical range of quality is substantially narrower than in niche, unmediated markets. The main challenge posed by content in social media sites is the fact that the distribution of quality has high variance from very high-quality items to low-quality, sometimes abusive, content.

Reach: Both industrial and social media technologies provide scale and are capable of reaching a global audience. Industrial media, however, typically use a centralized framework for organization, production, and dissemination, whereas social media

are by their very nature more decentralized, less hierarchical, and distinguished by multiple points of production and utility.

Frequency: Compared to industrial media, updating and reposting on social media platforms is easier, faster, and cheaper, and therefore practiced more frequently, resulting in fresher content.

Accessibility: The means of production for industrial media are typically government and/or corporate (privately owned) and are costly, whereas social media tools are generally available to the public at little or no cost.

Usability: Industrial media production typically requires specialized skills and training. Conversely, most social media production requires only modest reinterpretation of existing skills; in theory, anyone with access can operate the means of social media production.

Immediacy: The time lag between communications produced by industrial media can be long (weeks, months, or even years) compared to social media (which can be capable of virtually instantaneous responses).

Updatability: Industrial media, once created, cannot be altered (once a magazine article is printed and distributed, changes cannot be made to that same article), whereas social media can be altered almost instantaneously by comments or editing.

How Do People Use Social Media?

Not only are the numbers on social networking sites growing, but so is the degree to which they are engaged with the channel. Brogan and Bastone (2011) presented research results that stratify users according to how actively they use social media and tracked the evolution of these user segments over time. They listed six different engagement levels (Figure 7.15).

According to the research results, the online user community has been steadily migrating upward on this engagement hierarchy. The most notable change is among Inactives. Of the online population, 44 percent fell into this category in 2008. Two years later, more than half of those Inactives had jumped into social media in some form or another. "Now roughly 82% of the adult population online is in one of the upper categories," said Bastone. "Social media has truly reached a state of mass adoption" (Brogan and Bastone, 2011).

Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness. Social media analytics is rapidly

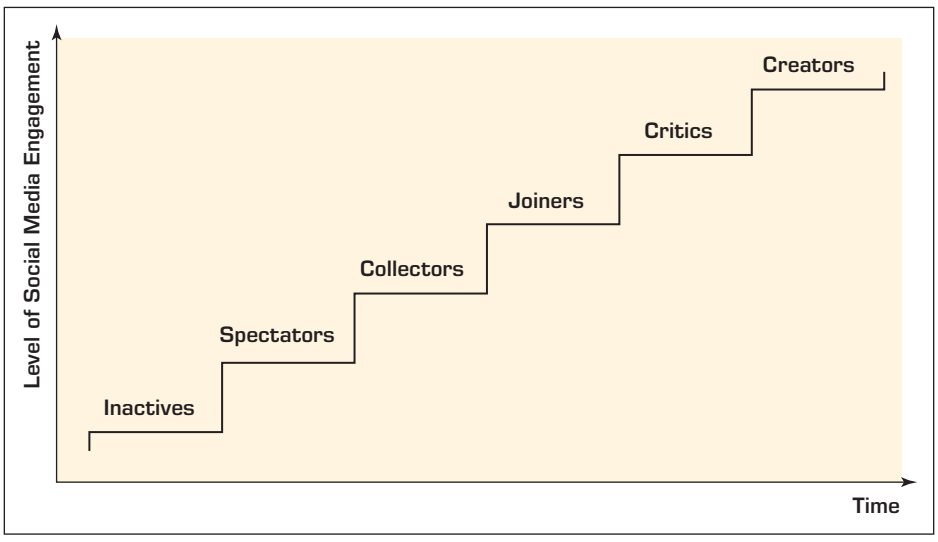


FIGURE 7.15 Evolution of Social Media User Engagement.

becoming a new force in organizations around the world, allowing them to reach out to and understand consumers as never before. In many companies, it is becoming the tool for integrated marketing and communications strategies.

The exponential growth of social media outlets from blogs, Facebook, and Twitter to LinkedIn and YouTube and of analytics tools that tap into these rich data sources offer organizations the chance to join a conversation with millions of customers around the globe every day. This ability is why nearly two-thirds of the 2,100 companies who participated in a recent survey by Harvard Business Review Analytic Services said they were either currently using social media channels or had social media plans in the works (Harvard Business Review, 2010). But many still say social media is an experiment, as they try to understand how to best use the different channels, gauge their effectiveness, and integrate social media into their strategy.

Measuring the Social Media Impact

For organizations, small or large, there is valuable insight hidden in all the user-generated content on social media sites. But how do you dig it out of dozens of review sites, thousands of blogs, millions of Facebook posts, and billions of tweets? Once you do that, how do you measure the impact of your efforts? These questions can be addressed by the analytics extension of the social media technologies. Once you decide on your goal for social media (what it is that you want to accomplish), a multitude of tools can help you get there. These analysis tools usually fall into three broad categories:

- ***Descriptive analytics:*** Uses simple statistics to identify activity characteristics and trends, such as how many followers you have, how many reviews were generated on Facebook, and which channels are being used most often.
- ***Social network analysis:*** Follows the links between friends, fans, and followers to identify connections of influence as well as the biggest sources of influence.
- ***Advanced analytics:*** Includes predictive analytics and text analytics that examine the *content* in online conversations to identify themes, sentiments, and connections that would not be revealed by casual surveillance.

Sophisticated tools and solutions to social media analytics use all three categories of analytics (i.e., descriptive, predictive, and prescriptive) in a somewhat progressive fashion.

Best Practices in Social Media Analytics

As an emerging tool, social media analytics is practiced by companies in a somewhat haphazard fashion. Because there are not well-established methodologies, everybody is trying to create their own by trial and error. What follows are some of the best field-tested practices for social media analytics proposed by Paine and Chaves (2012).

THINK OF MEASUREMENT AS A GUIDANCE SYSTEM, NOT A RATING SYSTEM Measurements are often used for punishment or rewards; they should not be. They should be about figuring out what the most effective tools and practices are, what needs to be discontinued because it does not work, and what needs to be done more because it does work very well. A good analytics system should tell you where you need to focus. Maybe all that emphasis on Facebook does not really matter because that is not where your audience is. Maybe they are all on Twitter, or vice versa. According to Paine and Chaves (2012), channel preference will not necessarily be intuitive: "We just worked with a hotel that had virtually no activity on Twitter for one brand but lots of Twitter activity for one of their higher brands." Without an accurate measurement tool, you would not know.

TRACK THE ELUSIVE SENTIMENT Customers want to take what they are hearing and learning from online conversations and act on it. The key is to be precise in extracting and tagging their intentions by measuring their sentiments. As we saw earlier in this chapter, text

analytic tools can categorize online content, uncover linked concepts, and reveal the sentiment in a conversation as "positive," "negative," or "neutral," based on the words people use. Ideally, you would like to be able to attribute sentiment to a specific product, service, and business unit. The more precise you can be in understanding the tone and perception that people express, the more actionable the information becomes because you are mitigating concerns about mixed polarity. A mixed-polarity phrase, such as "hotel in great location but bathroom was smelly," should not be tagged as "neutral" because you have positives and negatives offsetting each other. To be actionable, these types of phrases are to be treated separately; "bathroom was smelly" is something someone can own and improve on. One can classify and categorize these sentiments, look at trends over time, and see significant differences in the way people speak either positively or negatively about you. Furthermore, you can compare sentiment about your brand to your competitors.

CONTINUOUSLY IMPROVE THE ACCURACY OF TEXT ANALYSIS An industry-specific text analytics package will already know the vocabulary of your business. The system will have linguistic rules built into it, but it learns over time and gets better and better. Much as you would tune a statistical model as you have more data, better parameters, or new techniques to deliver better results, you would do the same thing with the NLP that goes into sentiment analysis. You set up rules, taxonomies, categorization, and meaning of words; watch what the results look like and then go back and do it again.

LOOK AT THE RIPPLE EFFECT It is one thing to be a great hit on a high-profile site, but that is only the start. There is a difference between a great hit that just sits there and goes away versus a great hit that is tweeted, retweeted, and picked up by influential bloggers. Analysis should show you which social media activities go "viral" and which quickly go dormant—and why.

LOOK BEYOND THE BRAND One of the biggest mistakes people make is to be concerned only about their brand. To successfully analyze and act on social media, people need to understand not just what is being said about their brand but also the broader conversation about the spectrum of issues surrounding their product or service, as well. Customers do not usually care about a firm's message or its brand; they care about themselves. Therefore, you should pay attention to what they are talking about, where they are talking, and where their interests are.

IDENTIFY YOUR MOST POWERFUL INFLUENCERS Organizations struggle to identify who has the most power in shaping public opinion. It turns out, their most important influencers are not necessarily the ones who advocate specifically for their brand; they are the ones who influence the whole realm of conversation about their topic. Organizations need to understand whether influencers are saying nice things, expressing support, or simply making observations or critiquing. What is the nature of their conversations? How is the organization's brand being positioned relative to the competition in that space?

LOOK CLOSELY AT THE ACCURACY OF ANALYTIC TOOLS USED Until recently, computer-based automated tools were not as accurate as humans for sifting through online content. Even now, accuracy varies depending on the media. For product review sites, hotel review sites, and Twitter, the accuracy can reach anywhere between 80 and 90 percent because the context is more boxed in. When an organization starts looking at blogs and discussion forums where the conversation is more wide ranging, the software can deliver 60 to 70 percent accuracy (Paine & Chaves, 2012). These figures will increase over time because the analytics tools are continually upgraded with new rules and improved algorithms to reflect field experience, new products, changing market conditions, and emerging patterns of speech.

INCORPORATE SOCIAL MEDIA INTELLIGENCE INTO PLANNING Once an organization has a big-picture perspective and detailed insight, it can begin to incorporate this information into

its planning cycle. But that is easier said than done. A quick audience poll revealed that very few people currently incorporate learning from online conversations into their planning cycles (Paine & Chaves, 2012). One way to achieve this is to find time-linked associations between social media metrics and other business activities or market events. Social media is typically either organically invoked or invoked by something an organization does; therefore, if it sees a spike in activity at some point in time, it wants to know what was behind that.

SECTION 7.10 REVIEW QUESTIONS

1. What is meant by social analytics? Why is it an important business topic?
2. What is a social network? What is the need for SNA?
3. What is social media? How does it relate to Web 2.0?
4. What is social media analytics? What are the reasons behind its increasing popularity?
5. How can you measure the impact of social media analytics?

Chapter Highlights

- Text mining is the discovery of knowledge from unstructured (mostly text-based) data sources. Because a great deal of information is in text form, text mining is one of the fastest-growing branches of the business intelligence field.
- Text mining applications are in virtually every area of business and government, including marketing, finance, healthcare, medicine, and homeland security.
- Text mining uses NLP to induce structure into the text collection and then uses data mining algorithms such as classification, clustering, association, and sequence discovery to extract knowledge from it.
- *Sentiment* can be defined as a settled opinion reflective of one's feelings.
- Sentiment analysis deals with differentiating between two classes, positive and negative.
- As a field of research, sentiment analysis is closely related to computational linguistics, NLP, and text mining.
- Sentiment analysis is trying to answer the question, "What do people feel about a certain topic?" by digging into opinions of many by using a variety of automated tools.
- VOC is an integral part of an analytic CRM and customer experience management systems and is often powered by sentiment analysis.
- VOM is about understanding aggregate opinions and trends at the market level.
- Polarity identification in sentiment analysis is accomplished either by using a lexicon as a reference library or by using a collection of training documents.
- WordNet is a popular general-purpose lexicon created at Princeton University.
- SentiWordNet is an extension of WordNet to be used for sentiment identification.
- Speech analytics is a growing field of science that allows users to analyze and extract information from both live and recorded conversations.
- *Web mining* can be defined as the discovery and analysis of interesting and useful information from the Web, about the Web, and usually using Web-based tools.
- Web mining can be viewed as consisting of three areas: content mining, structure mining, and usage mining.
- Web content mining refers to the automatic extraction of useful information from Web pages. It can be used to enhance search results produced by search engines.
- Web structure mining refers to generating interesting information from the links on Web pages.
- Web structure mining can also be used to identify the members of a specific community and perhaps even the roles of the members in the community.
- Web usage mining refers to developing useful information through analyzing Web server logs, user profiles, and transaction information.
- Text and Web mining are emerging as critical components of the next generation of business intelligence tools to enable organizations to compete successfully.
- A search engine is a software program that searches for documents (Internet sites or files) based on the keywords (individual words, multiword terms, or a complete sentence) users have provided that relate to the subject of their inquiry.
- SEO is the intentional activity of affecting the visibility of an e-commerce site or a Web site

in a search engine's natural (unpaid or organic) search results.
- VOC is a term generally used to describe the analytic process of capturing a customer's expectations, preferences, and aversions.
- Social analytics is the monitoring, analyzing, measuring, and interpreting of digital interactions and relationships of people, topics, ideas, and content.
- A social network is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships.
- Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques to better an organization's competitiveness.

Key Terms

association
authoritative pages
classification
clickstream analysis
clustering
corpus
deception detection
hubs
hyperlink-induced topic search (HITS)
natural language processing (NLP)
part-of-speech tagging
polarity identification
polyseme
search engine
sentiment analysis
SentiWordNet
singular value decomposition (SVD)
social media analytics
social network
spider
stemming
stop words
term–document matrix (TDM)
text mining
tokenizing
trend analysis
unstructured data
voice of the customer (VOC)
Web analytics
Web content mining
Web crawler
Web mining
Web structure mining
Web usage mining
WordNet

Questions for Discussion

1. Explain the relationship among data mining, text mining, and sentiment analysis.
2. In your own words, define *text mining*, and discuss its most popular applications.
3. What does it mean to induce structure into text-based data? Discuss the alternative ways of inducing structure into them.
4. What is the role of NLP in text mining? Discuss the capabilities and limitations of NLP in the context of text mining.
5. List and discuss three prominent application areas for text mining. What is the common theme among the three application areas you chose?
6. What is sentiment analysis? How does it relate to text mining?
7. What are the common challenges with which sentiment analysis deals?
8. What are the most popular application areas for sentiment analysis? Why?
9. What are the main steps in carrying out sentiment analysis projects?
10. What are the two common methods for polarity identification? Explain.
11. Discuss the differences and commonalities between text mining and Web mining.
12. In your own words, define Web mining, and discuss its importance.
13. What are the three main areas of Web mining? Discuss the differences and commonalities among these three areas.
14. What is a search engine? Why is it important for businesses?
15. What is SEO? Who benefits from it? How?
16. What is Web analytics? What are the metrics used in Web analytics?
17. Define *social analytics, social network*, and *social network analysis*. What are the relationships among them?
18. What is social media analytics? How is it done? Who does it? What comes out of it?

Exercises

Teradata University Network (TUN) and Other Hands-on Exercises

1. Visit **teradatauniversitynetwork.com**. Identify cases about text mining. Describe recent developments in the field. If you cannot find enough cases at the Teradata University Network Web site, broaden your search to other Web-based resources.
2. Go to **teradatauniversitynetwork.com** to locate white papers, Web seminars, and other materials related to text mining. Synthesize your findings into a short written report.

3. Go to **teradatauniversitynetwork.com** and find the case study named "eBay Analytics." Read the case carefully and extend your understanding of it by searching the Internet for additional information, and answer the case questions.
4. Go to **teradatauniversitynetwork.com** and find the sentiment analysis case named "How Do We Fix an App Like That?" Read the description, and follow the directions to download the data and the tool to carry out the exercise.
5. Visit **teradatauniversitynetwork.com**. Identify cases about Web mining. Describe recent developments in the field. If you cannot find enough cases at the Teradata University Network Web site, broaden your search to other Web-based resources.
6. Browse the Web and your library's digital databases to identify articles that make the linkage between text/Web mining and contemporary business intelligence systems.

Team Assignments and Role-Playing Projects

1. Examine how textual data can be captured automatically using Web-based technologies. Once captured, what are the potential patterns that you can extract from these unstructured data sources?
2. Interview administrators at your college or executives in your organization to determine how text mining and Web mining could assist them in their work. Write a proposal describing your findings. Include a preliminary cost–benefit analysis in your report.
3. Go to your library's online resources. Learn how to download attributes of a collection of literature (journal articles) in a specific topic. Download and process the data using a methodology similar to the one explained in Application Case 7.5.
4. Find a readily available sentiment text data set (see Technology Insights 7.2 for a list of popular data sets) and download it onto your computer. If you have an analytics tool that is capable of text mining, use that. If not, download RapidMiner (**http://rapid-i.com**) and install it. Also install the Text Analytics add-on for RapidMiner. Process the downloaded data using your text mining tool (i.e., convert the data into a structured form). Build models and assess the sentiment detection accuracy of several classification models (e.g., support vector machines, decision trees, neural networks, logistic regression). Write a detailed report in which you explain your findings and your experiences.
5. Examine how Web-based data can be captured automatically using the latest technologies. Once captured, what are the potential patterns that you can extract from these content-rich, mostly unstructured data sources?

Internet Exercises

1. Find recent cases of successful text mining and Web mining applications. Try text and Web mining software vendors and consultancy firms and look for cases or success stories. Prepare a report summarizing five new case studies.
2. Go to **statsoft.com**. Select Downloads, and download at least three white papers on applications. Which of these applications might have used the data/text/Web mining techniques discussed in this chapter?
3. Go to **sas.com**. Download at least three white papers on applications. Which of these applications might have used the data/text/Web mining techniques discussed in this chapter?
4. Go to **ibm.com**. Download at least three white papers on applications. Which of these applications might have used the data/text/Web mining techniques discussed in this chapter?
5. Go to **teradata.com**. Download at least three white papers on applications. Which of these applications might have used the data/text/Web mining techniques discussed in this chapter?
6. Go to **clarabridge.com**. Download at least three white papers on applications. Which of these applications might have used text mining in a creative way?
7. Go to **kdnuggets.com**. Explore the sections on applications as well as software. Find names of at least three additional packages for data mining and text mining.
8. Survey some Web mining tools and vendors. Identify some Web mining products and service providers that are not mentioned in this chapter.
9. Go to **attensity.com**. Download at least three white papers on Web analytics applications. Which of these applications might have used a combination of data/text/Web mining techniques?

References

Bond, C. F., & B. M. DePaulo. (2006). "Accuracy of Deception Judgments." *Personality and Social Psychology Reports, 10*(3), pp. 214–234.

Brogan, C., & J. Bastone. (2011). "Acting on Customer Intelligence from Social Media: The New Edge for Building Customer Loyalty and Your Brand." SAS white paper.

Chun, H. W., Y. Tsuruoka, J. D. Kim, R. Shiba, N. Nagata, & T. Hishiki. (2006). "Extraction of Gene-Disease Relations from MEDLINE Using Domain Dictionaries and Machine Learning." *Proceedings of the Eleventh Pacific Symposium on Biocomputing*, pp. 4–15.

Coussement, K., & D. Van Den Poel. (2008). "Improving Customer Complaint Management by Automatic Email Classification Using Linguistic Style Features as Predictors." *Decision Support Systems, 44*(4), pp. 870–882.

Coussement, K., & D. Van Den Poel. (2009). "Improving Customer Attrition Prediction by Integrating Emotions from Client/Company Interaction Emails and Evaluating Multiple Classifiers." *Expert Systems with Applications, 36*(3), pp. 6127–6134.

Cutts, M. (2006, February 4). "Ramping Up on International Webspam." **mattcutts.com/blog**.**mattcutts.com/blog/ramping-up-on-international-webspam** (accessed March 2013).

Delen, D., & M. Crossland. (2008). "Seeding the Survey and Analysis of Research Literature with Text Mining." *Expert Systems with Applications, 34*(3), pp. 1707–1720.

Esuli, A., & F. Sebastiani. (2006, May). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC, 6*, pp. 417–422.

Etzioni, O. (1996). "The World Wide Web: Quagmire or Gold Mine?" *Communications of the ACM, 39*(11), pp. 65–68.

EUROPOL. (2007). EUROPOL Work Program 2005. **statewatch.org/news/2006/apr/europol-work-programme-2005.pdf** (accessed October 2008).

Feldman, R., & J. Sanger. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Boston, MA: ABS Ventures.

Fuller, C. M., D. Biros, and D. Delen. (2008). "Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection." *Proceedings of the Forty-First Annual Hawaii International Conference on System Sciences (HICSS)*. Big Island, HI: IEEE Press, pp. 80–99.

Ghani, R., K. Probst, Y. Liu, M. Krema, and A. Fano. (2006). "Text Mining for Product Attribute Extraction." *SIGKDD Explorations, 8*(1), pp. 41–48.

Goodman, A. (2005). "Search Engine Showdown: Black Hats Versus White Hats at SES. SearchEngineWatch." **searchenginewatch.com/article/2066090/Search-Engine-Showdown-Black-Hats-vs.-White-Hats-at-SES** (accessed February 2013).

Han, J., & M. Kamber. (2006). *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann.

Harvard Business Review. (2010). "The New Conversation: Taking Social Media from Talk to Action." A SAS–Sponsored Research Report by Harvard Business Review Analytic Services. **sas.com/resources/whitepaper/wp_23348.pdf** (accessed March 2013).

Kaplan, A. M., & M. Haenlein. (2010). "Users of the World, Unite! The Challenges and Opportunities of Social Media." *Business Horizons, 53*(1), pp. 59–68.

Kim, S. M., & E. Hovy. (2004, August). "Determining the Sentiment of Opinions." *Proceedings of the Twentieth International Conference on Computational Linguistics*, p. 1367.

Kleinberg, J. (1999). "Authoritative Sources in a Hyperlinked Environment." *Journal of the ACM, 46*(5), pp. 604–632.

Lin, J., & D. Demner-Fushman. (2005). "Bag of Words" Is Not Enough for Strength of Evidence Classification." *AMIA Annual Symposium Proceedings*, pp. 1031–1032. **pubmedcentral.nih.gov/articlerender.fcgi?artid=1560897**.

Liu, B., M. Hu, & J. Cheng. (2005, May). "Opinion Observer: Analyzing and Comparing Opinions on the Web." *Proceedings of the Fourth International Conference on World Wide Web*, pp. 342–351.

Mahgoub, H., D. Rösner, N. Ismail, and F. Torkey. (2008). "A Text Mining Technique Using Association Rules Extraction." *International Journal of Computational Intelligence, 4*(1), pp. 21–28.

Manning, C. D., & H. Schutze. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McKnight, W. (2005, January 1). "Text Data Mining in Business Intelligence." *Information Management Magazine*. **information-management.com/issues/20050101/1016487-1.html** (accessed May 22, 2009).

Mejova, Y. (2009). "Sentiment Analysis: An Overview." Comprehensive exam paper. **http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf** (accessed February 2013).

Miller, T. W. (2005). *Data and Text Mining: A Business: Applications Approach*. Upper Saddle River, NJ: Prentice Hall.

Morgan, N., G. Jones, & A. Hodges. (2010). "The Complete Guide to Social Media from the Social Media Guys." **thesocialmediaguys.co.uk/wp-content/uploads/downloads/2011/03/CompleteGuidetoSocialMedia.pdf** (accessed February 2013).

Nakov, P., A. Schwartz, B. Wolf, and M. A. Hearst. (2005). "Supporting Annotation Layers for Natural Language Processing." *Proceedings of the ACL*, Interactive Poster and Demonstration Sessions. Ann Arbor, MI: Association for Computational Linguistics, pp. 65–68.

Paine, K. D., & M. Chaves. (2012). "Social Media Metrics." SAS white paper. **sas.com/resources/whitepaper/wp_19861.pdf** (accessed February 2013).

Pang, B., & L. Lee. (2008). *OPINION Mining and Sentiment Analysis*. Hanover, MA: Now Publishers; available at **http://books.google.com**.

Ramage, D., D. Hall, R. Nallapati, & C. D. Manning. (2009, August). "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pp. 248–256.

Schmidt, L.-H. (1996). "Commonness Across Cultures." In A. N. Balslev (ed.), *Cross-Cultural Conversation: Initiation* (pp. 119–132). New York: Oxford University Press.

Scott, W. R., & G. F. Davis. (2003). "Networks in and Around Organizations." *Organizations and Organizing*. Upper Saddle River: NJ: Pearson Prentice Hall.

Shatkay, H., A. Höglund, S. Brady, T. Blum, P. Dönnes, and O. Kohlbacher. (2007). "SherLoc: High-Accuracy Prediction of Protein Subcellular Localization by Integrating Text and Protein Sequence Data." *Bioinformatics, 23*(11), pp. 1410–1415.

Snyder, B., & R. Barzilay. (2007, April). "Multiple Aspect Ranking Using the Good Grief Algorithm." *HLT-NAACL*, pp. 300–307.

Strapparava, C., & A. Valitutti. (2004, May). "WordNet Affect: An Affective Extension of WordNet." *LREC, 4*, pp. 1083–1086.

The Westover Group. (2013). "20 Key Web Analytics Metrics and How to Use Them." **http://www.thewestovergroup.com** (accessed February 2013).

Thomas, M., B. Pang, & L. Lee. (2006, July). "Get Out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts." In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 327–335.

Weng, S. S., & C. K. Liu. (2004). "Using Text Classification and Multiple Concepts to Answer E-Mails." *Expert Systems with Applications, 26*(4), pp. 529–543.

PART III

Prescriptive Analytics and Big Data

CHAPTER

8

Prescriptive Analytics: Optimization and Simulation

LEARNING OBJECTIVES

- Understand the applications of prescriptive analytics techniques in combination with reporting and predictive analytics
- Understand the basic concepts of analytical decision modeling
- Understand the concepts of analytical models for selected decision problems, including linear programming and simulation models for decision support
- Describe how spreadsheets can be used for analytical modeling and solutions
- Explain the basic concepts of optimization and when to use them
- Describe how to structure a linear programming model
- Explain what is meant by sensitivity analysis, what-if analysis, and goal seeking
- Understand the concepts and applications of different types of simulation
- Understand potential applications of discrete event simulation

This chapter extends the analytics applications beyond reporting and predictive analytics. It includes coverage of selected techniques that can be employed in combination with predictive models to help support decision making. We focus on techniques that can be implemented relatively easily using either spreadsheet tools or by using stand-alone software tools. Of course, there is much additional detail to be learned about management science models, but the objective of this chapter is to simply illustrate what is possible and how it has been implemented in real settings.

We present this material with a note of caution: Modeling can be a difficult topic and is as much an art as it is a science. The purpose of this chapter is not necessarily for you to *master the topics* of modeling and analysis. Rather, the material is geared toward *gaining familiarity* with the important concepts as they relate to prescriptive analytics and their use in decision making. It is important to recognize that the modeling we discuss here is only cursorily related to the concepts of data modeling. You should not confuse the two. We walk through some basic concepts and definitions of decision modeling. We next introduce the idea of modeling directly in spreadsheets. We then discuss the structure and application of two successful time-proven models

and methodologies: linear programming and discrete event simulation. As noted earlier, one could take multiple courses just in these two topics, but our goal is to give you a sense of what is possible. This chapter includes the following sections:

8.1 OPENING VIGNETTE: School District of Philadelphia Uses Prescriptive Analytics to Find Optimal Solution for Awarding Bus Route Contracts

BACKGROUND

Selecting the best vendors to work with is a laborious yet important task for companies and government organizations. After a vendor submits a proposal for a specific task through a bidding process, the company or organization evaluates the proposal and makes a decision on which vendor is best suited for their needs. Typically, governments are required to use a bidding process to select one or more vendors. The School District of Philadelphia was in search of private bus vendors to outsource some of their bus routes. The district owned a few school buses, but needed more to serve their student population. They wanted to use their own school buses for 30 to 40% of the routes, and outsource the rest of the routes to these private vendors. Charles Lowitz, the fiscal coordinator for the transportation office, was tasked with determining how to maximize the return on investment and refine the way routes were awarded to various vendors.

Historically, the process of deciding which bus vendor contracts to award given the budget and time constraints was laborious as it was done manually by hand. In addition, the different variables and factors that had to be taken into account added to the complexity. The vendors were evaluated based on five variables: cost, capabilities, reliance, financial stability, and business acumen. Each vendor submitted a proposal with a different price for different routes. Some vendors specified a minimum number of routes, and if that minimum wasn't met, their cost would increase. Lowitz needed to figure out how to combine the information from each proposal to determine which bus route to award to which vendor to meet all the route requirements at the least cost for the district.

SOLUTION

Lowitz initially looked for software that he could use in conjunction with his contract model in Excel. He began using the Premium Solver Platform from Frontline Systems, Inc., which allowed him to find the most beneficial vendors for the district from a financial and operational standpoint. He created an optimization model that took into account the aforementioned variables associated with each vendor. The model included binary integer variables (yes/no) for each of the routes to be awarded to the bidders

who proposed to serve a specific route at a specific cost. This amounted to about 1,600 yes/no variables. The model also included constraints indicating that each route was to be awarded to one vendor, and of course, each route had to be serviced. Other constraints specified the minimum number of routes a vendor would accept and a few other details. All such constraints can be written as equations and entered in an integer linear programming model. Such models can be formulated and solved through many software tools, but using Microsoft Excel makes it easier to understand the model. Frontline Systems' Solver software is built into Microsoft Excel to solve smaller problems for free. A larger version can be purchased to solve larger and more complex models. That is what Lowitz used.

BENEFITS

In addition to determining how many of the vendors should be awarded contracts, the model helped develop the size of each of the contracts. The size of the contracts varied from one vendor getting four routes to another receiving 97 routes. Ultimately, the School District of Philadelphia was able to create a plan with an optimized number of bus company vendors using Excel instead of a manual handwritten process. By using the Premium Solver Platform analytic tools to create an optimization model with the different variables, the district saved both time and money.

➧ QUESTIONS FOR THE OPENING VIGNETTE

1. What decision was being made in this vignette?
2. What data (descriptive and or predictive) might one need to make the best allocations in this scenario?
3. What other costs or constraints might you have to consider in awarding contracts for such routes?
4. Which other situations might be appropriate for applications of such models?

WHAT CAN WE LEARN FROM THIS OPENING VIGNETTE?

Most organizations face the problem of making decisions where one has to select from multiple options. Each option has a cost and capability associated with it. The goal of such models is to select the combination of options that meet all the requirements and yet optimizes the costs. Prescriptive analytics particularly apply to the problem of such decisions. And tools such as built-in or Premium Solver for Excel make it easy to apply such techniques.

Source: Based on "Optimizing Vendor Contract Awards Gets an A+," **http://www.solver.com/news/optimizing-vendor-contract-awards-gets**, 2016 (accessed Sept 2018).

8.2 MODEL-BASED DECISION MAKING

As the preceding vignette indicates, making decisions using some kind of analytical model is what we call prescriptive analytics. In the last several chapters we have learned the value and the process of knowing the history of what has been going on and use that information to also predict what is likely to happen. However, we go through that exercise to determine what we should do next. This might entail deciding which customers are likely to buy from us and making an offer or giving a price point that will maximize the likelihood that they would buy and our profit would be optimized. Conversely, it might involve being able to predict which customer is likely to go somewhere else and

making a promotion offer to retain them as a customer and optimize our value. We may need to make decisions on awarding contracts to our vendors to make sure all our needs are covered and the costs are minimized. We could be facing a situation of deciding which prospective customers should receive what promotional campaign material so that our cost of promotion is not outrageous, and we maximize the response rate while managing within a budget. We may be deciding how much to pay for different paid search keywords to maximize the return on investment of our advertising budget. In another setting, we may have to study the history of our customers' arrival patterns and use that information to predict future arrival rates, and apply that to schedule an appropriate number of store employees to maximize customer responses and optimize our labor costs. We could be deciding where to locate our warehouses based on our analysis and prediction of demand for our products and the supply chain costs. We could be setting daily delivery routes on the basis of product volumes to be delivered at various locations and the delivery costs and vehicle availability. One can find hundreds of examples of situations where data-based decisions are valuable. Indeed, the biggest opportunity for the growing analytics profession is the ability to use descriptive and predictive insights to help a decision maker make better decisions. Although there are situations where one can use experience and intuition to make decisions, it is more likely that a decision supported by a model will help a decision maker make better decisions. In addition, it also provides decision makers with justification for what they are recommending. Thus prescriptive analytics has emerged as the next frontier for analytics. It essentially involves using an analytical model to help guide a decision maker in making a decision, or automating the decision process so that a model can make recommendations or decisions. Because the focus of prescriptive analytics is on making recommendations or making decisions, some call this category of analytics decision analytics.

INFORMS publications, such as *Interfaces, ORMS Today*, and *Analytics* magazine, all include stories that illustrate successful applications of decision models in real settings. This chapter includes many examples of such prescriptive analytic applications. Applying models to real-world situations can save millions of dollars or generate millions of dollars in revenue. Christiansen et al. (2009) describe the applications of such models in shipping company operations using TurboRouter, a decision support system (DSS) for ship routing and scheduling. They claim that over the course of a 3-week period, a company used this model to better utilize its fleet, generating additional profit of $1−2 million in such a short time. We provide another example of a model application in Application Case 8.1 that illustrates a sports application.

Application Case 8.1 Canadian Football League Optimizes Game Schedule

Canadian Football League (CFL) is Canada's equivalent of the U.S. National Football League (NFL). It had a challenge of organizing 81 football games for 9 teams over a period of 5 months optimally while stabilizing matching priorities for sales revenue, television ratings, and the team rest days. Other considerations include organizing games over different time zones and the main rivalry games to be held on major public holidays. For any league, a robust schedule is a driving force for a variety of business collaborations, such as coordinating with broadcasting channels and organizing ground ticket sales. If the schedule is not optimized, it would directly hamper the promotions thus resulting in a huge loss of revenue and bad channel ratings. CFL used to create match schedules manually and

(*Continued*)

Application Case 8.1 (Continued)

hence had to figure out finer ways to improve their schedules, taking all the constraints into account. They had tried to work with a consultant to build a comprehensive model for scheduling, but the implementation remained a challenge. The League decided to tackle the issue with the Solver available within Microsoft Excel. Some of the matching priorities to be balanced while optimizing the schedule were:

1. Sales Revenue—Setting a schedule with matches and time slots to those clubs that generate more revenue.
2. Channel Ratings—Setting a schedule with games that would improve channel ratings for the broadcasting company.
3. Team Rest Days—Setting a schedule with the two teams playing against each other having enough rest days.

The league decided to improve the match schedules by giving the player rest days as a higher priority, followed by sales revenue and channel scores for the broadcasting company. This is mainly because the sales revenue and channel scores are a byproduct of team players' performance on the field, which is directly related to the rest days taken by the teams.

Methodology/Solution

Initially, organizing schedules was a huge task to perform on Excel through the built-in Solver feature. Frontline systems provided a premium version for Solver which allowed the model size to grow from about 200 decisions to 8,000 decisions. The League had to even add in more industry-specific constraints such as telecasting across different time zones, double header games cannot be overlapped, and arch rival games to be scheduled on Labor Day. Added limitations were never simple until the Frontline Systems consultants stepped up to help CFL turn this nonlinear problem into a linear problem. The linear programming "engine" got the model running. Premium Solver software turned out to be of great help to get an improved schedule.

Results/Benefits

Using the optimized schedule would lead to increased revenue through higher ticket sales and higher TV scores for the broadcasting channels. This was achieved because the tool was able to support added constraints of the vendors with great ease. The optimized schedule pleased most of the league's stakeholders. This is a repetitive process, but those match schedules were CFL's most advanced season match schedules to date.

Questions for Discussion

1. List three ways in which Solver-based scheduling of games could result in more revenue as compared to the manual scheduling.
2. In what other ways can CFL leverage the Solver software to expand and enhance their other business operations?
3. What other considerations could be important in scheduling such games?

What Can We Learn from This Application Case?

By using the Solver add-in for Excel, the CFL made better decisions in scheduling their games by taking stakeholders and industry constraints into consideration, leading to revenue generation and good channel ratings. Thus, an optimized schedule, a purview of prescriptive analytics, derived significant value. According to the case study, the modeler, Mr Trevor Hardy, was an expert Excel user, but not an expert in modeling. However, the ease of use of Excel permitted him to develop a practical application of prescriptive analytics.

Compiled from "Canadian Football League Uses Frontline Solvers to Optimize Scheduling in 2016." Solver, September 7 2016, **www.solver.com/news/canadian-football-league-uses-frontline-solvers-optimize-scheduling-2016** (accessed September 2018); Kostuk, Kent J., and Keith A. Willoughby. "A Decision Support System for Scheduling the Canadian Football League." *Interfaces,* vol. 42, no. 3, 2012, pp. 286–295; Dilkina, Bistra N., and William S. Havens. The U.S. National Football League Scheduling Problem. Intelligent Systems Lab, **www.cs.cornell.edu/~bistra/papers/NFLsched1.pdf** (accessed September 2018).

Prescriptive Analytics Model Examples

Modeling is a key element for prescriptive analytics. In the examples mentioned earlier in the introduction and application cases, one has to employ a mathematical model to be able to recommend a decision for any realistic problem. For example, deciding which customers (among potentially millions) will receive what offer so as to maximize the overall response value but staying within a budget is not something you can do manually. Building a probability-based response maximization model with the budget as a constraint would give us the information we are seeking. Depending on the problem we are addressing, there are many classes of models, and there are often many specialized techniques for solving each one. We will learn about two different modeling methods in this chapter. Most universities have multiple courses that cover these topics under titles such as Operations Research, Management Science, Decision Support Systems, and Simulation that can help you build more expertise in these topics. Because prescriptive analytics typically involves the application of mathematical models, sometimes the term *data science* is more commonly associated with the application of such mathematical models. Before we learn about mathematical modeling support in prescriptive analytics, let us understand some modeling issues first.

Identification of the Problem and Environmental Analysis

No decision is made in a vacuum. It is important to analyze the scope of the domain and the forces and dynamics of the environment. A decision maker needs to identify the organizational culture and the corporate decision-making processes (e.g., who makes decisions, degree of centralization). It is entirely possible that environmental factors have created the current problem. This can formally be called **environmental scanning and analysis**, which is the monitoring, scanning, and interpretation of collected information. Business intelligence/business analytics (BI/BA) tools can help identify problems by scanning for them. The problem must be understood, and everyone involved should share the same frame of understanding because the problem will ultimately be represented by the model in one form or another. Otherwise, the model will not help the decision maker.

VARIABLE IDENTIFICATION Identification of a model's variables (e.g., decision, result, uncontrollable) is critical, as are the relationships among the variables. Influence diagrams, which are graphical models of mathematical models, can facilitate the identification process. A more general form of an influence diagram, a cognitive map, can help a decision maker develop a better understanding of a problem, especially of variables and their interactions.

FORECASTING (PREDICTIVE ANALYTICS) As we have noted previously, an important prerequisite of prescriptive analytics is knowing what has happened and what is likely to happen. This form of predictive analytics is essential for construction and manipulating models because when a decision is implemented, the results usually occur in the future. There is no point in running a what-if (sensitivity) analysis on the past because decisions made then have no impact on the future. Online commerce and communication has created an immense need for **forecasting** and an abundance of available information for performing it. These activities occur quickly, yet information about such purchases is gathered and should be analyzed to produce forecasts. Part of the analysis involves simply predicting demand; however, forecasting models can use product life-cycle needs and information about the marketplace and consumers to analyze the entire situation, ideally leading to additional sales of products and services.

We describe an effective example of such forecasting and its use in decision making at Ingram Micro in Application Case 8.2.

Application Case 8.2

Ingram Micro Uses Business Intelligence Applications to Make Pricing Decisions

Ingram Micro is the world's largest two-tier distributor of technology products. In a two-tier distribution system, a company purchases products from manufacturers and sells them to retailers who in turn sell these products to the end users. For example, one can purchase a Microsoft Office 365 package from Ingram rather than purchasing it directly from Microsoft. Ingram has partnerships with Best Buy, Buffalo, Google, Honeywell, Libratone, and Sharper Image. The company delivers its products to 200,000 solution providers across the world and thus has a large volume of transaction data. Ingram wanted to use insights from this data to identify cross-selling opportunities and determine prices to offer to specific customers in conjunction with product bundles. This required setting up a business intelligence center (BIC) to compile and analyze the data. In setting up the BIC, Ingram faced various issues.

1. Ingram faced several issues in their data-capture process such as a lack of loss data, ensuring the accuracy of end-user information, and linking quotes to orders.
2. Ingram faced technical issues in implementing a customer relationship management (CRM) system capable enough to handle its operations around the world.
3. They faced resistance to the idea of demand pricing (determining price based on demand of product).

Methodology/Solution

Ingram explored communicating directly with its customers (resellers) using e-mail and offered them discounts on the purchase of supporting technologies related to the products being ordered. They identified these opportunities through segmented market-basket analysis and developed the following business intelligence applications that helped in determining optimized prices. Ingram developed a new price optimization tool known as IMPRIME, which is capable of setting data-driven prices and providing data-driven negotiation guidance. IMPRIME sets an optimized price for each level of the product hierarchy (i.e., customer level, vendor-customer level, customer-segment level, and vendor-customer segment level). It does so by taking into account the trade-off between the demand signal and pricing at that level.

The company also developed a digital marketing platform known as Intelligence INGRAM. This platform utilizes predictive lead scoring (PLS), which selects end users to target with specific marketing programs. PLS is their system to score predictive leads for companies that have no direct relation with end users. Intelligence INGRAM is used to run white space programs, which encourage a reseller to purchase related products by offering discounts. For example, if a reseller purchases a server from INGRAM, then INGRAM offers a discount on disk storage units as both products are required to work together. Similarly, Intelligence INGRAM is used to run growth incentive campaigns (offering cash rewards to resellers on exceeding quarterly spend goals) and cross-sell campaigns (e-mailing the end users about the products that are related to their recently purchased product).

Results/Benefits

Profit generated by using IMPRIME is measured using a lift measurement methodology. This methodology compares periods before and after changing the prices and compares test groups versus control groups. Lift measurement is done on average daily sales, gross margin, and machine margin. The use of IMPRIME led to a $757 million growth in revenue and a $18.8 million increase in gross profits.

Questions for Discussion

1. What were the main challenges faced by Ingram Micro in developing a BIC?
2. List all the business intelligence solutions developed by Ingram to optimize the prices of their products and to profile their customers.
3. What benefits did Ingram receive after using the newly developed BI applications?

What Can We Learn from This Application Case?

By first building a BIC, a company begins to better understand its product lines, its customers, and their purchasing patterns. This insight is derived from what we call descriptive and predictive analytics. Further value from this is derived through price optimization, a purview of prescriptive analytics.

Sources: R. Mookherjee, J. Martineau, L. Xu, M. Gullo, K. Zhou, A. Hazlewood, X. Zhang, F. Griarte, & N. Li. (2016). "End-to-End Predictive Analytics and Optimization in Ingram Micro's Two-Tier Distribution Business." *Interfaces, 46*(1), 49–73; **ingrammicro-commerce.com**, "CUSTOMERS," **https://www.ingrammicro-commerce.com/customers/** (accessed July 2016).

Model Categories

Table 8.1 classifies some decision models into seven groups and lists several representative techniques for each category. Each technique can be applied to either a **static** or a **dynamic model**, which can be constructed under assumed environments of certainty, uncertainty, or risk. To expedite model construction, we can use special decision analysis systems that have modeling languages and capabilities embedded in them. These include spreadsheets, data mining systems, online analytic processing (OLAP) systems, and modeling languages that help an analyst build a model. We will introduce one of these systems later in the chapter.

MODEL MANAGEMENT Models, like data, must be managed to maintain their integrity, and thus their applicability. Such management is done with the aid of model-based management systems, which are analogous to database management systems (DBMS).

KNOWLEDGE-BASED MODELING DSS uses mostly quantitative models, whereas expert systems use qualitative, knowledge-based models in their applications. Some knowledge is necessary to construct solvable (and therefore usable) models. Many of the predictive

TABLE 8.1 Categories of Models

Category	Process and Objective	Representative Techniques
Optimization of problems with few alternatives	Find the best solution from a small number of alternatives	Decision tables, decision trees, analytic hierarchy process
Optimization via algorithm	Find the best solution from a large number of alternatives, using a step-by-step improvement process	Linear and other mathematical programming models, network models
Optimization via an analytic formula	Find the best solution in one step, using a formula	Some inventory models
Simulation	Find a good enough solution or the best among the alternatives checked, using experimentation	Several types of simulation
Heuristics	Find a good enough solution, using rules	Heuristic programming, expert systems
Predictive models	Predict the future for a given scenario	Forecasting models, Markov analysis
Other models	Solve a what-if case, using a formula	Financial modeling, waiting lines

analytics techniques, such as classification and clustering, can be used in building knowledge-based models.

CURRENT TRENDS IN MODELING One recent trend in modeling involves the development of model libraries and solution technique libraries. Some of these codes can be run directly on the owner's Web server for free, and others can be downloaded and run on a local computer. The availability of these codes means that powerful optimization and simulation packages are available to decision makers who may have only experienced these tools from the perspective of classroom problems. For example, the Mathematics and Computer Science Division at Argonne National Laboratory (Argonne, Illinois) maintains the NEOS Server for Optimization at **https://neos-server.org/neos/index.html**. You can find links to other sites by clicking the Resources link at informs.org, the Web site of the Institute for Operations Research and the Management Sciences (INFORMS). A wealth of modeling and solution information is available from INFORMS. The Web site for one of INFORMS' publications, *OR/MS Today,* at **http://www.orms-today.org/ormsmain.shtml** includes links to many categories of modeling software. We will learn about some of these shortly.

There is a clear trend toward developing and using cloud-based tools and software to access and even run software to perform modeling, optimization, simulation, and so on. This has, in many ways, simplified the application of many models to real-world problems. However, to use models and solution techniques effectively, it is necessary to truly gain experience through developing and solving simple ones. This aspect is often overlooked. Organizations that have key analysts who understand how to apply models indeed apply them very effectively. This is most notably occurring in the revenue management area, which has moved from the province of airlines, hotels, and automobile rentals to retail, insurance, entertainment, and many other areas. CRM also uses models, but they are often transparent to the user. With management models, the amount of data and model sizes are quite large, necessitating the use of data warehouses to supply the data and parallel computing hardware to obtain solutions in a reasonable time frame.

There is a continuing trend toward making analytics models completely transparent to the decision maker. For example, **multidimensional analysis (modeling)** involves data analysis in several dimensions. In multidimensional analysis (modeling), data are generally shown in a spreadsheet format, with which most decision makers are familiar. Many decision makers accustomed to slicing and dicing data cubes are now using OLAP systems that access data warehouses. Although these methods may make modeling palatable, they also eliminate many important and applicable model classes from consideration, and they eliminate some important and subtle solution interpretation aspects. Modeling involves much more than data analysis with trend lines and establishing relationships with statistical methods.

There is also a trend to build a model of a model to help in its analysis. An **influence diagram** is a graphical representation of a model; that is, a model of a model. Some influence diagram software packages are capable of generating and solving the resultant model.

SECTION 8.2 REVIEW QUESTIONS

1. List three lessons learned from modeling.
2. List and describe the major issues in modeling.
3. What are the major types of models used in DSS?
4. Why are models not used in industry as frequently as they should or could be?
5. What are the current trends in modeling?

8.3 STRUCTURE OF MATHEMATICAL MODELS FOR DECISION SUPPORT

In the following sections, we present the topics of analytical mathematical models (e.g., mathematical, financial, and engineering). These include the components and the structure of models.

The Components of Decision Support Mathematical Models

All **quantitative models** are typically made up of four basic components (see Figure 8.1): result (or outcome) variables, decision variables, uncontrollable variables (and/or parameters), and intermediate result variables. Mathematical relationships link these components together. In nonquantitative models, the relationships are symbolic or qualitative. The results of decisions are determined based on the decision made (i.e., the values of the decision variables), the factors that cannot be controlled by the decision maker (in the environment), and the relationships among the variables. The modeling process involves identifying the variables and relationships among them. Solving a model determines the values of these and the result variable(s).

RESULT (OUTCOME) VARIABLES **Result (outcome) variables** reflect the level of effectiveness of a system; that is, they indicate how well the system performs or attains its goal(s). These variables are outputs. Examples of result variables are shown in Table 8.2. Result variables are considered *dependent variables*. Intermediate result variables are sometimes used in modeling to identify intermediate outcomes. In the case of a dependent variable, another event must occur first before the event described by the variable can occur. Result variables depend on the occurrence of the decision variables and the uncontrollable variables.

DECISION VARIABLES **Decision variables** describe alternative courses of action. The decision maker controls the decision variables. For example, for an investment problem, the amount to invest in bonds is a decision variable. In a scheduling problem, the decision variables are people, times, and schedules. Other examples are listed in Table 8.2.

UNCONTROLLABLE VARIABLES, OR PARAMETERS In any decision-making situation, there are factors that affect the result variables but are not under the control of the decision maker. Either these factors can be fixed, in which case they are called **uncontrollable variables**, or **parameters**, or they can vary, in which case they are called *variables*. Examples of factors are the prime interest rate, a city's building code, tax regulations, and utilities costs. Most of these factors are uncontrollable because they are in and determined by elements of the system environment in which the decision maker works. Some of

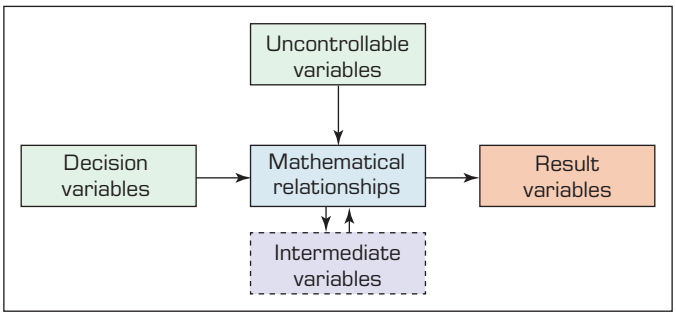


FIGURE 8.1 The General Structure of a Quantitative Model.

TABLE 8.2 Examples of Components of Models

Area	Decision Variables	Result Variables	Uncontrollable Variables and Parameters
Financial investment	Investment alternatives and amounts	Total profit, risk Rate of return on investment (ROI) Earnings per share Liquidity level	Inflation rate Prime rate Competition
Marketing	Advertising budget Where to advertise	Market share Customer satisfaction	Customer's income Competitor's actions
Manufacturing	What and how much to produce Inventory levels Compensation programs	Total cost Quality level Employee satisfaction	Machine capacity Technology Materials prices
Accounting	Use of computers Audit schedule	Data processing cost Error rate	Computer technology Tax rates Legal requirements
Transportation	Shipments schedule Use of smart cards	Total transport cost Payment float time	Delivery distance Regulations
Services	Staffing levels	Customer satisfaction	Demand for services

these variables limit the decision maker and therefore form what are called *constraints* of the problem.

INTERMEDIATE RESULT VARIABLES **Intermediate result variables** reflect intermediate outcomes in mathematical models. For example, in determining machine scheduling, spoilage is an intermediate result variable, and total profit is the result variable (i.e., spoilage is one determinant of total profit). Another example is employee salaries. This constitutes a decision variable for management: It determines employee satisfaction (i.e., intermediate outcome), which, in turn, determines the productivity level (i.e., final result).

The Structure of Mathematical Models

The components of a quantitative model are linked by mathematical (algebraic) expressions—equations or inequalities.

A very simple financial model is

$$P = R - C$$

where P = profit, R = revenue, and C = cost. This equation describes the relationship among the variables. Another well-known financial model is the simple present-value cash flow model, where P = present value, F = a future single payment in dollars, i = interest rate (percentage), and n = number of years. With this model, it is possible to determine the present value of a payment of \$100,000 to be made 5 years from today, at a 10% (0.1) interest rate, as follows:

$$P = 100{,}000/(1 + 0.1)^5 = 62{,}092$$

We present more interesting and complex mathematical models in the following sections.

SECTION 8.3 REVIEW QUESTIONS

1. What is a decision variable?
2. List and briefly discuss the major components of a quantitative model.
3. Explain the role of intermediate result variables.

8.4 CERTAINTY, UNCERTAINTY, AND RISK

The[1] decision-making process involves evaluating and comparing alternatives. During this process, it is necessary to predict the future outcome of each proposed alternative. Decision situations are often classified on the basis of what the decision maker knows (or believes) about the forecasted results. We customarily classify this knowledge into three categories (see Figure 8.2), ranging from complete knowledge to complete ignorance:

- Certainty
- Uncertainty
- Risk

When we develop models, any of these conditions can occur, and different kinds of models are appropriate for each case. Next, we discuss both the basic definitions of these terms and some important modeling issues for each condition.

Decision Making under Certainty

In decision making under **certainty**, it is *assumed* that complete knowledge is available so that the decision maker knows exactly what the outcome of *each course of action* will be (as in a deterministic environment). It may not be true that the outcomes are 100% known, nor is it necessary to really evaluate *all* the outcomes, but often this assumption simplifies the model and makes it tractable. The decision maker is viewed as a perfect predictor of the future because it is assumed that there is only one outcome for each alternative. For example, the alternative of investing in U.S. Treasury bills is one for which there is complete availability of information about the future return on investment if it is held to maturity. A situation involving decision making under certainty occurs most often with structured problems and short time horizons (up to 1 year). Certainty models are relatively easy to develop and solve, and they can yield optimal solutions. Many financial models are constructed under assumed certainty, even though the market is anything but 100% certain.

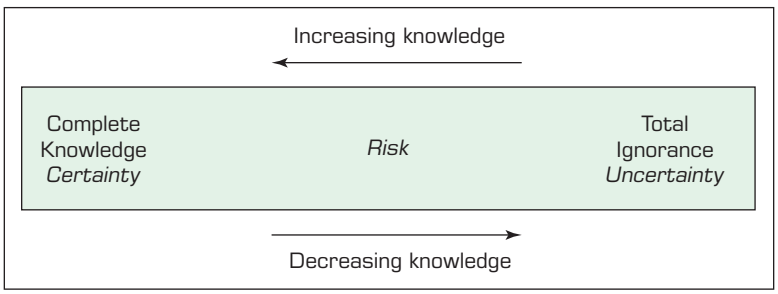


FIGURE 8.2 The Zones of Decision Making.

[1]Some parts of the original versions of these sections were adapted from Turban and Meredith (1994).

Decision Making under Uncertainty

In decision making under **uncertainty**, the decision maker considers situations in which several outcomes are possible for each course of action. In contrast to the risk situation, in this case, the decision maker does not know, or cannot estimate, the probability of occurrence of the possible outcomes. Decision making under uncertainty is more difficult than decision making under certainty because there is insufficient information. Modeling of such situations involves assessment of the decision maker's (or the organization's) attitude toward risk.

Managers attempt to avoid uncertainty as much as possible, even to the point of assuming it away. Instead of dealing with uncertainty, they attempt to obtain more information so that the problem can be treated under certainty (because it can be "almost" certain) or under calculated (i.e., assumed) risk. If more information is not available, the problem must be treated under a condition of uncertainty, which is less definitive than the other categories.

Decision Making under Risk (Risk Analysis)

A decision made under **risk**[2] (also known as a *probabilistic* or *stochastic* decision-making situation) is one in which the decision maker must consider several possible outcomes for each alternative, each with a given probability of occurrence. The long-run probabilities that the given outcomes will occur are assumed to be known or can be estimated. Under these assumptions, the decision maker can assess the degree of risk associated with each alternative (called *calculated* risk). Most major business decisions are made under assumed risk. **Risk analysis** (i.e., calculated risk) is a decision-making method that analyzes the risk (based on assumed known probabilities) associated with different alternatives. Risk analysis can be performed by calculating the expected value of each alternative and selecting the one with the best expected value. Application Case 8.3 illustrates one application to reduce uncertainty.

Application Case 8.3

American Airlines Uses Should-Cost Modeling to Assess the Uncertainty of Bids for Shipment Routes

American Airlines, Inc. (AA) is one of the world's largest airlines. Its core business is passenger transportation, but it has other vital ancillary functions that include full-truckload (FTL) freight shipment of maintenance equipment and in-flight shipment of passenger service items that could add up to over $1 billion in inventory at any given time. AA receives numerous bids from suppliers in response to requests for quotes (RFQs) for inventories. AA's RFQs could total over 500 in any given year. Bid quotes vary significantly as a result of the large number of bids and resultant complex bidding process. Sometimes, a single contract bid could deviate by about 200%. As a result of the complex process, it is common to either overpay or underpay suppliers for their services. To this end, AA wanted a should-cost model that would streamline and assess bid quotes from suppliers to choose bid quotes that were fair to both them and their suppliers.

Methodology/Solution

To determine fair cost for supplier products and services, three steps were taken:

1. Primary (e.g., interviews) and secondary (e.g., Internet) sources were scouted for base-case

[2]Our definitions of the terms *risk* and *uncertainty* were formulated by F. H. Knight of the University of Chicago in 1933. Other, comparable definitions also are in use.

and range data that would inform cost variables that affect an FTL bid.
2. Cost variables were chosen so that they were mutually exclusive and collectively exhaustive.
3. The DPL decision analysis software was used to model the uncertainty.

Furthermore, Extended Swanson-Megill approximation was used to model the probability distribution of the most sensitive cost variables used. This was done to account for the high variability in the bids in the initial model.

Results/Benefits

A pilot test was done on an RFQ that attracted bids from six FTL carriers. Out of the six bids presented, five were within three standard deviations from the mean, whereas one was considered an outlier. Subsequently, AA used the should-cost FTL model on more than 20 RFQs to determine what a fair and accurate cost of goods and services should be. It is expected that this model will help in reducing the risk of either overpaying or underpaying its suppliers.

Questions for Discussion

1. Besides reducing the risk of overpaying or underpaying suppliers, what are some other benefits AA would derive from its "should-be" model?
2. Can you think of other domains besides air transportation where such a model could be used?
3. Discuss other possible methods with which AA could have solved its bid overpayment and underpayment problem.

Source: Based on Bailey, M. J., Snapp, J., Yetur, S., Stonebraker, J. S., Edwards, S. A., Davis, A., & Cox, R. (2011). Practice summaries: American Airlines uses should-cost modeling to assess the uncertainty of bids for its full-truckload shipment routes. *Interfaces*, *41*(2), 194–196.

SECTION 8.4 REVIEW QUESTIONS

1. Define what it means to perform decision making under assumed certainty, risk, and uncertainty.
2. How can decision-making problems under assumed certainty be handled?
3. How can decision-making problems under assumed uncertainty be handled?
4. How can decision-making problems under assumed risk be handled?

8.5 DECISION MODELING WITH SPREADSHEETS

Models can be developed and implemented in a variety of programming languages and systems. We focus primarily on *spreadsheets* (with their add-ins), modeling languages, and transparent data analysis tools. With their strength and flexibility, spreadsheet packages were quickly recognized as easy-to-use implementation software for the development of a wide range of applications in business, engineering, mathematics, and science. Spreadsheets include extensive statistical, forecasting, and other modeling and database management capabilities, functions, and routines. As spreadsheet packages evolved, add-ins were developed for structuring and solving specific model classes. Among the add-in packages, many were developed for DSS development. These DSS-related add-ins include Solver (Frontline Systems Inc., **solver.com**) and What's*Best!* (a version of Lindo, from Lindo Systems, Inc., **lindo.com**) for performing linear and nonlinear optimization; Braincel (Jurik Research Software, Inc., **jurikres.com**) and NeuralTools (Palisade Corp., **palisade.com**) for artificial neural networks; Evolver (Palisade Corp.) for genetic algorithms; and @RISK (Palisade Corp.) for performing simulation studies. Comparable add-ins are available for free or at a very low cost. (Conduct a Web search to find them; new ones are added to the marketplace on a regular basis.)

The spreadsheet is clearly the most popular *end-user modeling tool* because it incorporates many powerful financial, statistical, mathematical, and other functions. Spreadsheets can perform model solution tasks such as linear programming and regression analysis. The spreadsheet has evolved into an important tool for analysis, planning, and modeling (see Farasyn, Perkoz, & Van de Velde, 2008; Hurley & Balez, 2008; Ovchinnikov & Milner, 2008). Application Cases 8.4 and 8.5 describe interesting applications of spreadsheet-based models in a nonprofit setting.

Application Case 8.4

Pennsylvania Adoption Exchange Uses Spreadsheet Model to Better Match Children with Families

The Pennsylvania Adoption Exchange (PAE) was established in 1979 by the State of Pennsylvania to help county and nonprofit agencies find prospective families for orphan children who had not been adopted due to age or special needs. The PAE keeps detailed records about children and preferences of families who may adopt them. The exchange looks for families for the children across all 67 counties of Pennsylvania.

The Pennsylvania Statewide Adoption and Permanency Network is responsible for finding permanent homes for orphans. If after a few attempts the network fails to place a child with a family, they then get help from the PAE. The PAE uses an automated assessment tool to match children to families. This tool gives matching recommendations by calculating a score between 0 and 100% for a child on 78 pairs of the child's attribute values and family preferences. For some years now, the PAE has struggled to give adoption match recommendations to caseworkers for children. They are finding it difficult to manage a vast database of children collected over time for all 67 counties. The basic search algorithm produced match recommendations that were proving unfruitful for caseworkers. As a result, the number of children who have not been adopted has increased significantly, and there is a growing urgency to find families for these orphans.

Methodology/Solution

The PAE started collecting information about the orphans and families through online surveys that include a new set of questions. These questions collect information about hobbies of the child, child–caseworker preferences for families, and preference of the age range of children by families. The PAE and consultants created a spreadsheet matching tool that included additional features compared to the previously used automated tool. In this model, caseworkers can specify the weight of the attributes for selecting a family for a child. For example, if a family had a narrow set of preferences regarding gender, age, and race, then those factors can receive a higher weight. Also, caseworkers can give preference about the family's county of residence, as community relationship is an important factor for a child. Using this tool, the matching committee can compare a child and family on each attribute, thus making a more accurate match decision between a family and a child.

Results/Benefits

Since the PAE started using the new spreadsheet model for matching a family with a child, they have been able to make better matching decisions. As a result, the percentage of children getting a permanent home has increased.

This short case is one of the many examples of using spreadsheets as a decision support tool. By creating a simple scoring system for a family's desire and a child's attribute, a better matching system is produced so that fewer rejections are reported on either side.

Questions for Discussion

1. What were the challenges faced by PAE while making adoption matching decisions?
2. What features of the new spreadsheet tool helped PAE solve their issues of matching a family with a child?

Source: Based on Slaugh, V. W., Akan, M., Kesten, O., & Unver, M. U. (2016). The Pennsylvania Adoption Exchange improves its matching process. *Interfaces*, *46*(2), 133–154.

Application Case 8.5

Metro Meals on Wheels Treasure Valley Uses Excel to Find Optimal Delivery Routes

Meals on Wheels Association of America (now Meals on Wheels America) is a not-for-profit organization that delivers approximately one million meals to homes of older people in need across the United States. Metro Meals on Wheels Treasure Valley is a local branch of Meals on Wheels America operating in Idaho. This branch has a team of volunteer drivers that drive their personal vehicles each day to deliver meals to 800 clients along 21 routes and cover an area of 2,745 square kilometers.

The Meals on Wheels Treasure Valley organization was facing many issues. First, they were looking to minimize the delivery time as the cooked food was temperature sensitive and could perish easily. They wanted to deliver the cooked food within 90 minutes after a driver left for the delivery. Second, the scheduling process was very time consuming. Two employees spent much of their time developing scheduled routes for delivery. A route coordinator determined the stops according to the number of meal recipients for a given day. After determining the stops, the coordinator made a sequence of stops that minimized the travel time of volunteers. This routing schedule was then entered into an online tool to determine turn-by-turn driving instructions for drivers. The whole process of manually deciding routes was taking a lot of extra time. Metro Meals on Wheels wanted a routing tool that could improve their delivery system and generate routing solutions for both one-way and round-trip directions for delivering meals. Those who drive regularly could deliver the warmers or coolers the next day. Others who drive only occasionally would need to come back to the kitchen to drop off the warmers/coolers.

Methodology/Solution

To solve the routing problem, a spreadsheet-based tool was developed. This tool had an interface to easily input information about the recipient such as his/her name, meal requirements, and delivery address. This information needed to be filled in the spreadsheet for each stop in the route. Next, Excel's Visual Basic for Applications functionality was used to access a developer's networking map application programming interface (API) called MapQuest. This API was used to create a travel matrix that calculated time and distance needed for delivery of the meal. This tool gave time and distance information for 5,000 location pairs a day without any cost.

When the program starts, the MapQuest API first validates the entered addresses of meal recipients. Then the program uses the API to retrieve driving distance, estimated driving time, and turn-by-turn instructions for driving between all stops in the route. The tool can then find the optimal route for up to 30 stops within a feasible time limit.

Results/Benefits

As a result of using this tool, the total annual driving distance decreased by 10,000 miles, while travel time was reduced by 530 hours. Metro Meals on Wheels Treasure Valley saved $5,800 in 2015, based on an estimated savings rate of $0.58 per mile (for a midsize sedan). This tool also reduced the time spent on route planning for meal deliveries. Other benefits included increased volunteer satisfaction and more retention of volunteers.

Questions for Discussion

1. What were the challenges faced by Metro Meals on Wheels Treasure Valley related to meal delivery before adoption of the spreadsheet-based tool?
2. Explain the design of the spreadsheet-based model.
3. What are the intangible benefits of using the Excel-based model to Metro Meals on Wheels?

Source: Based on Manikas, A. S., Kroes, J. R., & Gattiker, T. F. (2016). Metro Meals on Wheels Treasure Valley employs a low-cost routing tool to improve deliveries. *Interfaces*, *46*(2), 154–167.

Other important spreadsheet features include what-if analysis, goal seeking, data management, and programmability (i.e., macros). With a spreadsheet, it is easy to change a cell's value and immediately see the result. Goal seeking is performed by indicating a target cell, its desired value, and a changing cell. Extensive database management can be performed with small data sets, or parts of a database can be imported for analysis (which

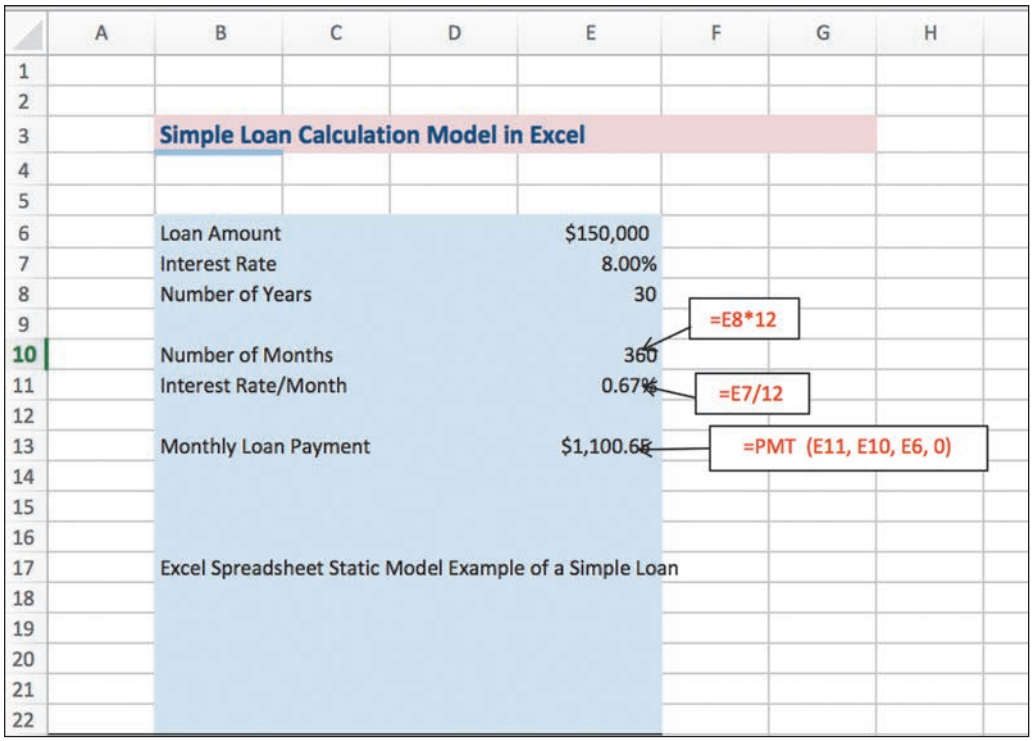


FIGURE 8.3 **Excel Spreadsheet Static Model Example of a Simple Loan Calculation of Monthly Payments.**

is essentially how OLAP works with multidimensional data cubes; in fact, most OLAP systems have the look and feel of advanced spreadsheet software after the data are loaded). Templates, macros, and other tools enhance the productivity of building DSS.

Most spreadsheet packages provide fairly seamless integration because they read and write common file structures and easily interface with databases and other tools. Microsoft Excel is the most popular spreadsheet package. In Figure 8.3, we show a simple loan calculation model in which the boxes on the spreadsheet describe the contents of the cells, which contain formulas. A change in the interest rate in cell E7 is immediately reflected in the monthly payment in cell E13. The results can be observed and analyzed immediately. If we require a specific monthly payment, we can use goal seeking to determine an appropriate interest rate or loan amount.

Static or dynamic models can be built in a spreadsheet. For example, the monthly loan calculation spreadsheet shown in Figure 8.3 is static. Although the problem affects the borrower over time, the model indicates a single month's performance, which is replicated. A dynamic model, in contrast, represents behavior over time. The loan calculations in the spreadsheet shown in Figure 8.4 indicate the effect of prepayment on the principal over time. Risk analysis can be incorporated into spreadsheets by using built-in random-number generators to develop simulation models (see the next chapter).

Spreadsheet applications for models are reported regularly. We will learn how to use a spreadsheet-based optimization model in the next section.

SECTION 8.5 REVIEW QUESTIONS

1. What is a spreadsheet?
2. What is a spreadsheet add-in? How can add-ins help in DSS creation and use?
3. Explain why a spreadsheet is so conducive to the development of DSS.

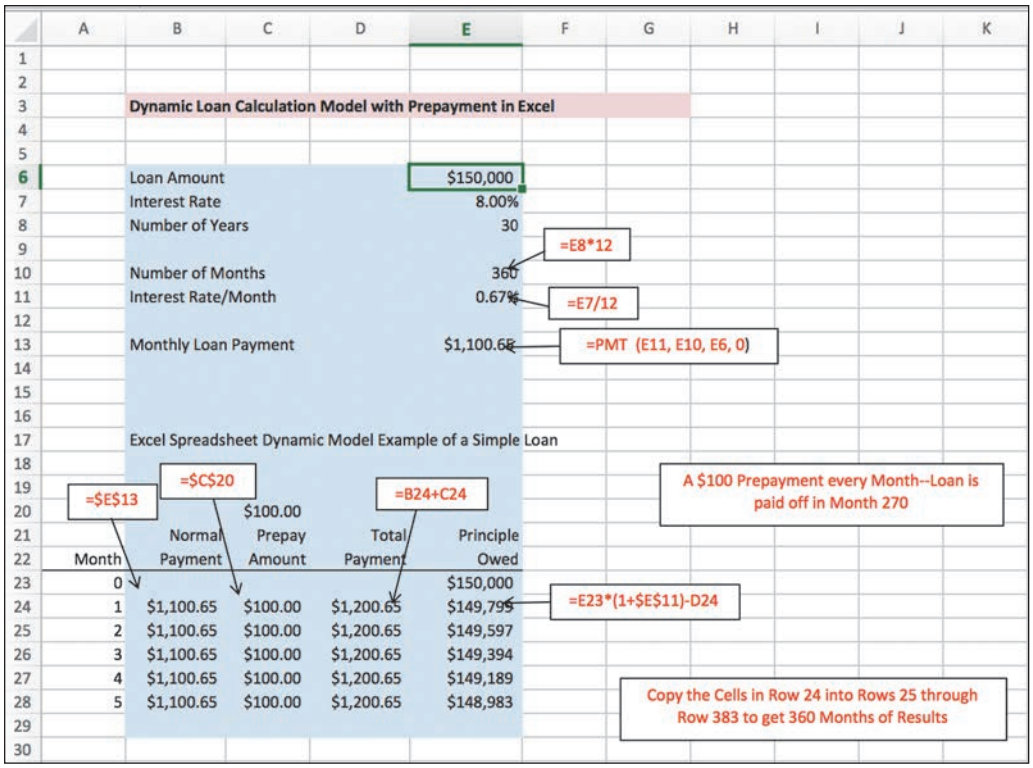


FIGURE 8.4 **Excel Spreadsheet Dynamic Model Example of a Simple Loan Calculation of Monthly Payments and the Effects of Prepayment.**

8.6 MATHEMATICAL PROGRAMMING OPTIMIZATION

Mathematical programming is a family of tools designed to help solve managerial problems in which the decision maker must allocate scarce resources among competing activities to optimize a measurable goal. For example, the distribution of machine time (the resource) among various products (the activities) is a typical allocation problem. **Linear programming (LP)** is the best-known technique in a family of optimization tools called *mathematical programming*; in LP, all relationships among the variables are linear. It is used extensively in DSS (see Application Case 8.6). LP models have many important applications in practice. These include supply chain management, product mix decisions, routing, and so on. Special forms of the models can be used for specific applications. For example, Application Case 8.6 describes a spreadsheet model that was used to create a schedule for physicians.

LP allocation problems usually display the following characteristics:

- A limited quantity of economic resources is available for allocation.
- The resources are used in the production of products or services.
- There are two or more ways in which the resources can be used. Each is called a *solution* or a *program*.
- Each activity (product or service) in which the resources are used yields a return in terms of the stated goal.
- The allocation is usually restricted by several limitations and requirements, called *constraints*.

Application Case 8.6

Mixed-Integer Programming Model Helps the University of Tennessee Medical Center with Scheduling Physicians

Regional Neonatal Associates is a nine-physician group working for the Neonatal Intensive Care Unit (NICU) at the University of Tennessee Medical Center in Knoxville, Tennessee. The group also serves two local hospitals in the Knoxville area for emergency purposes. For many years, one member of the group would schedule physicians manually. However, as his retirement approached, there was a need for a more automatic system to schedule physicians. The physicians wanted this system to balance their workload, as the previous schedules did not properly balance workload among them. In addition, the schedule needed to ensure there would be 24−7 NICU coverage by the physicians, and if possible, accommodate individual preferences of physicians for shift types. To address this problem, the physicians contacted the faculty of Management Science at the University of Tennessee.

The problem of scheduling physicians to shifts was characterized by constraints based on workload and lifestyle choices. The first step in solving the scheduling issue was to group shifts according to their types (day and night). The next step was determining constraints for the problem. The model needed to cover a nine-week period with nine physicians, with two physicians working weekdays and one physician overnight and on weekends. In addition, one physician had to be assigned exclusively for 24−7 coverage to the two local hospitals. Other obvious constraints also needed to be considered. For example, a day shift could not be assigned to a physician just after a night shift.

Methodology/Solution

The problem was formulated by creating a binary, mixed-integer optimization model. The first model divided workload equally among the nine physicians. But it could not assign an equal number of day and night shifts among them. This created a question of fair distribution. In addition, the physicians had differing opinions of the assigned workload. Six physicians wanted a schedule in which an equal number of day and night shifts would be assigned to each physician in the nine-week schedule, while the others wanted a schedule based on individual preference of shifts. To satisfy requirements of both groups of physicians, a new model was formed and named the Hybrid Preference Scheduling Model (HPSM). For satisfying the equality requirement of six physicians, the model first calculated one week's workload and divided it for nine weeks for them. This way, the work was divided equally for all six physicians. The workload for the three remaining physicians was distributed in the nine-week schedule according to their preference. The resulting schedule was reviewed by the physicians and they found the schedule more acceptable.

Results/Benefits

The HPSM method accommodated both the equality and individual preference requirements of the physicians. In addition, the schedules from this model provided better rest times for the physicians compared to the previous manual schedules, and vacation requests could also be accommodated in the schedules. The HPSM model can solve similar scheduling problems demanding relative preferences among shift types.

Techniques such as mixed-integer programming models can build optimal schedules and help in operations. These techniques have been used in large organizations for a long time. Now it is possible to implement such prescriptive analytic models in spreadsheets and other easily available software.

Questions for Discussion

1. What was the issue faced by the Regional Neonatal Associates group?
2. How did the HPSM model solve all of the physician's requirements?

Source: Adapted from Bowers, M. R., Noon, C. E., Wu, W., & Bass, J. K. (2016). Neonatal physician scheduling at the University of Tennessee Medical Center. *Interfaces*, *46*(2), 168–182.

The LP allocation model is based on the following rational economic assumptions:

- Returns from different allocations can be compared; that is, they can be measured by a common unit (e.g., dollars, utility).
- The return from any allocation is independent of other allocations.
- The total return is the sum of the returns yielded by the different activities.
- All data are known with certainty.
- The resources are to be used in the most economical manner.

Allocation problems typically have a large number of possible solutions. Depending on the underlying assumptions, the number of solutions can be either infinite or finite. Usually, different solutions yield different rewards. Of the available solutions, at least one is the best, in the sense that the degree of goal attainment associated with it is the highest (i.e., the total reward is maximized). This is called an **optimal solution**, and it can be found by using a special algorithm.

Linear Programming Model

Every LP model is composed of *decision variables* (whose values are unknown and are searched for), an *objective function* (a linear mathematical function that relates the decision variables to the goal, measures goal attainment, and is to be optimized), *objective function coefficients* (unit profit or cost coefficients indicating the contribution to the objective of one unit of a decision variable), *constraints* (expressed in the form of linear inequalities or equalities that limit resources and/or requirements; these relate the variables through linear relationships), *capacities* (which describe the upper and sometimes lower limits on the constraints and variables), and *input/output (technology) coefficients* (which indicate resource utilization for a decision variable).

Let us look at an example. MBI Corporation, which manufactures special-purpose computers, needs to make a decision: How many computers should it produce next month at the Boston plant? MBI is considering two types of computers: the CC-7, which requires 300 days of labor and $10,000 in materials, and the CC-8, which requires 500 days of labor and $15,000 in materials. The profit contribution of each CC-7 is $8,000, whereas that of each CC-8 is $12,000. The plant has a capacity of 200,000 working days per month, and the material budget is $8 million per month. Marketing requires that at least 100 units of the CC-7 and at least 200 units of the CC-8 be produced each month. The problem is to maximize the company's profits by determining how many units of the CC-7 and how many units of the CC-8 should be produced each month. Note that in a real-world environment, it could possibly take months to obtain the data in the problem statement, and

TECHNOLOGY INSIGHTS 8.1 Linear Programming

LP is perhaps the best-known optimization model. It deals with the optimal allocation of resources among competing activities. The allocation problem is represented by the model described here.

The problem is to find the values of the decision variables X_1, X_2, and so on, such that the value of the result variable Z is maximized, subject to a set of linear constraints that express the technology, market conditions, and other uncontrollable variables. The mathematical relationships are all linear equations and inequalities. Theoretically, any allocation problem of this type has an infinite number of possible solutions. Using special mathematical procedures, the LP approach applies a unique computerized search procedure that finds the best solution(s) in a matter of seconds. Furthermore, the solution approach provides automatic sensitivity analysis.

while gathering the data the decision maker would no doubt uncover facts about how to structure the model to be solved. Web-based tools for gathering data can help.

Modeling in LP: An Example

A standard LP model can be developed for the MBI Corporation problem just described. As discussed in Technology Insights 8.1, the LP model has three components: decision variables, result variables, and uncontrollable variables (constraints).
The decision variables are as follows:

$$X_1 = \text{units of CC}-7 \text{ to be produced}$$
$$X_2 = \text{units of CC}-8 \text{ to be produced}$$

The result variable is as follows:

$$\text{Total profit} = Z$$

The objective is to maximize total profit:

$$Z = 8{,}000X_1 + 12{,}000X_2$$

The uncontrollable variables (constraints) are as follows:

$$\text{Labor constraint: } 300X_1 + 500X_2 \leq 200{,}000 \text{ (in days)}$$
$$\text{Budget constraint: } 10{,}000X_1 + 15{,}000X_2 \leq 8{,}000{,}000 \text{ (in dollars)}$$
$$\text{Marketing requirement for CC}-7\text{: } X_1 \geq 100 \text{ (in units)}$$
$$\text{Marketing requirement for CC}-8\text{: } X_2 \geq 200 \text{ (in units)}$$

This information is summarized in Figure 8.5.

The model also has a fourth, hidden component. Every LP model has some internal intermediate variables that are not explicitly stated. The labor and budget constraints may each have some slack in them when the left-hand side is strictly less than the right-hand side. This slack is represented internally by slack variables that indicate excess resources available. The marketing requirement constraints may each have some surplus in them when the left-hand side is strictly greater than the right-hand side. This surplus is represented internally by surplus variables indicating that there is some room to adjust the right-hand sides of these constraints. These slack and surplus variables are intermediate. They can be of great value to a decision maker because LP solution methods use them in establishing sensitivity parameters for economic what-if analyses.

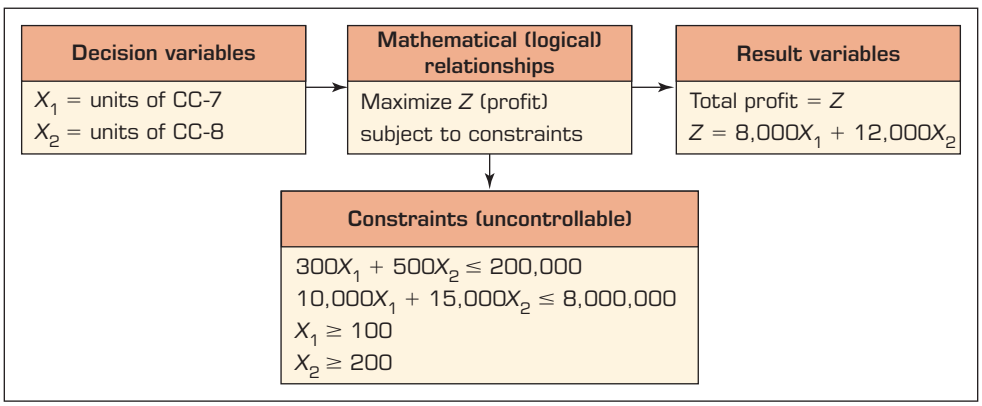


FIGURE 8.5 Mathematical Model of a Product-Mix Example.

The product-mix model has an infinite number of possible solutions. Assuming that a production plan is not restricted to whole numbers—which is a reasonable assumption in a monthly production plan—we want a solution that maximizes total profit: an optimal solution. Fortunately, Excel comes with the add-in Solver, which can readily obtain an optimal (best) solution to this problem. Although the location of Solver add-in has moved from one version of Excel to another, it is still available as a free add-in. Look for it under the Data tab and on the Analysis ribbon. If it is not there, you should be able to enable it by going to Excel's Options Menu and selecting Add-ins.

We enter these data directly into an Excel spreadsheet, activate Solver, and identify the goal (by setting Target Cell equal to Max), decision variables (by setting By Changing Cells), and constraints (by ensuring that Total Consumed elements is less than or equal to Limit for the first two rows and is greater than or equal to Limit for the third and fourth rows). Cells C7 and D7 constitute the decision variable cells. Results in these cells will be filled after running the Solver Add-in. Target Cell is Cell E7, which is also the result variable, representing a product of decision variable cells and their per unit profit coefficients (in Cells C8 and D8). Note that all the numbers have been divided by 1,000 to make it easier to type (except the decision variables). Rows 9–12 describe the constraints of the problem: the constraints on labor capacity, budget, and the desired minimum production of the two products X_1 and X_2. Columns C and D define the coefficients of these constraints. Column E includes the formulae that multiply the decision variables (Cells C7 and D7) with their respective coefficients in each row. Column F defines the right-hand side value of these constraints. Excel's matrix multiplication capabilities (e.g., SUMPRODUCT function) can be used to develop such row and column multiplications easily.

After the model's calculations have been set up in Excel, it is time to invoke the Solver Add-in. Clicking on the Solver Add-in (again under the Analysis group under Data Tab) opens a dialog box (window) that lets you specify the cells or ranges that define the objective function cell, decision/changing variables (cells), and the constraints. Also, in Options, we select the solution method (usually Simplex LP), and then we solve the problem. Next, we select all three reports—Answer, Sensitivity, and Limits—to obtain an optimal solution of $X_1 = 333.33$, $X_2 = 200$, Profit = \$5,066,667, as shown in Figure 8.6. Solver produces three useful reports about the solution. Try it. Solver now also includes the ability to solve nonlinear programming problems and integer programming problems by using other solution methods available within it.

The following example was created by Professor Rick Wilson of Oklahoma State University to further illustrate the power of spreadsheet modeling for decision support.

The table in Figure 8.7 describes some hypothetical data and attributes of nine "swing states" for the 2016 election. Attributes of the nine states include their number of electoral votes, two regional descriptors (note that three states are classified as neither North nor South), and an estimated "influence function," which relates to increased candidate support per unit of campaign financial investment in that state.

For instance, influence function F1 shows that for every financial unit invested in that state, there will be a total of a 10-unit increase in voter support (let units stay general here), made up of an increase in young men support by 3 units, old men support by 1 unit, and young and old women each by 3 units.

The campaign has 1,050 financial units to invest in the nine states. It must invest at least 5% in each state of the total overall invested, but no more than 25% of the overall total invested can be in any one state. All 1,050 units do not have to be invested (your model must correctly deal with this).

The campaign has some other restrictions as well. From a financial investment standpoint, the West states (in total) must have campaign investments at levels that are at least 60% of the total invested in East states. In terms of people influenced, the decision to allocate financial investments to states must lead to at least 9,200 total people influenced.

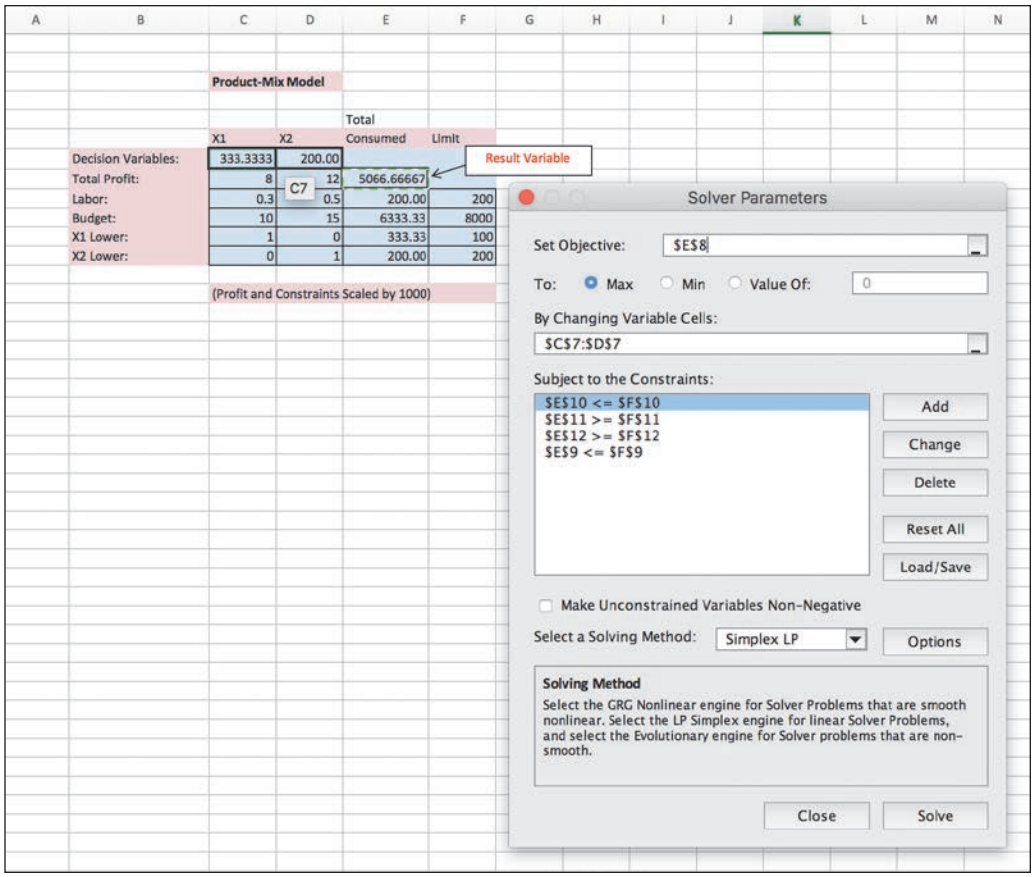


FIGURE 8.6 Excel Solver Solution to the Product-Mix Example.

	A	B	C	D	E	F	G	H
1								
2				Electoral			Influence	
3			State	Votes	W/E	N/S	Function	
4			NV	6	West		F1	
5			CO	9	West		F2	
6			IA	6	West	North	F3	
7			WI	10	West	North	F1	
8			OH	18	East	North	F2	
9			VA	13	East	South	F2	
10			NC	15	East	South	F1	
11			FL	29	East	South	F3	
12			NH	4	East		F3	
13								
14		F1	Young	Old				
15		Men	3	1	4			
16		Women	3	3	6			
17			6	4	10	Total		
18								
19		F2	Young	Old				
20		Men	1.5	2.5	4			
21		Women	2.5	1	3.5			
22			4	3.5	7.5	Total		
23								
24		F3	Young	Old				
25		Men	2.5	2.5	5			
26		Women	1	2	3			
27			3.5	4.5	8	Total		
28								

FIGURE 8.7 Data for Election Resource Allocation Example.

Overall, the total number of females influenced must be greater than or equal to the total number of males influenced. Also, at least 46% of all people influenced must be "old."

Our task is to create an appropriate integer programming model that determines the optimal integer (i.e., whole number) allocation of financial units to states that maximizes the sum of the products of the electoral votes times units invested subject to the other aforementioned restrictions. (Thus, indirectly, this model is giving preference to states with higher numbers of electoral votes.) Note that for ease of implementation by the campaign staff, all decisions for allocation in the model should lead to integer values.

The three aspects of the models can be categorized based on the following questions that they answer:

1. **What do we control?** The amount invested in advertisements across the nine states, Nevada, Colorado, Iowa, Wisconsin, Ohio, Virginia, North Carolina, Florida, and New Hampshire, which are represented by the nine decision variables, NV, CO, IA, WI OH, VA, NC, FL, and NH.
2. **What do we want to achieve?** We want to maximize the total number of electoral votes gains. We know the value of each electoral vote in each state (EV), so this amounts to EV*Investments aggregated over the nine states, that is,

$$\text{Max } (6NV + 9CO + 6IA + 10WI + 18OH + 13VA + 15NC + 29FL + 4NH)$$

3. **What constrains us?** Following are the constraints as given in the problem description:

   a. No more than 1,050 financial units to invest into, that is, NV + CO + IA + WI + OH + VA + NC + FL + NH <= 1,050.

   b. Invest at least 5% of the total in each state, that is,

$$NV >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$CO >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$IA >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$WI >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$OH >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$VA >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$NC >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$FL >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$NH >= 0.05\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

   We can implement these nine constraints in a variety of ways using Excel.

   c. Invest no more than 25% of the total in each state.

   As with (b) we need nine individual constraints again because we do not know how much of the 1,050 we will invest. We must write the constraints in "general" terms.

$$NV <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$CO <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$IA <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$WI <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$OH <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$VA <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$NC <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$FL <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$
$$NH <= 0.25\ (NV + CO + IA + WI + OH + VA + NC + FL + NH)$$

d. Western states must have investment levels that are at least 60% of the Eastern states.

West States = NV + CO + IA + WI

East States = OH + VA + NC + FL + NH

So, (NV + CO + IA + WI) >= 0.60 (OH + VA + NC + FL + NH). Again, we can implement this constraint in a variety of ways using Excel.

e. Influence at least 9,200 total people, that is,

(10NV + 7.5CO + 8IA + 10WI + 7.5OH + 7.5VA + 10NC
+ 8FL + 8 NH) >= 9,200

f. Influence at least as many females as males. This requires transition of influence functions.

F1 = 6 women influenced, F2 = 3.5 women

F3 = 3 women influenced

F1 = 4 men influenced, F2 = 4 men

F3 = 5 men influenced

So, implementing females >= males, we get:

(6NV + 3.5CO + 3IA + 6WI + 3.5OH + 3.5VA + 6NC + 3FL + 3NH) >=
(4NV + 4CO + 5IA + 4WI + 4OH + 4VA + 4NC + 5FL + 5NH)

As before, we can implement this in Excel in a couple of different ways.

g. At least 46% of all people influenced must be old.

All people influenced were on the left-hand side of the constraint (e). So, old people influenced would be:

(4NV + 3.5CO + 4.5IA + 4WI + 3.5OH + 3.5VA + 4NC + 4.5FL + 4.5NH)

This would be set >= 0.46* the left-hand side of constraint (e). (10NV + 7.5CO + 8IA + 10WI + 7.5OH + 7.5VA + 10NC + 8FL + 8NH), which would give a right-hand side of (0.46NV + 3.45CO + 3.68IA + 4.6WI + 3.45OH + 3.45VA + 4.6NC + 3.68FL + 3.68NH)

This is the last constraint other than to force all variables to be integers.

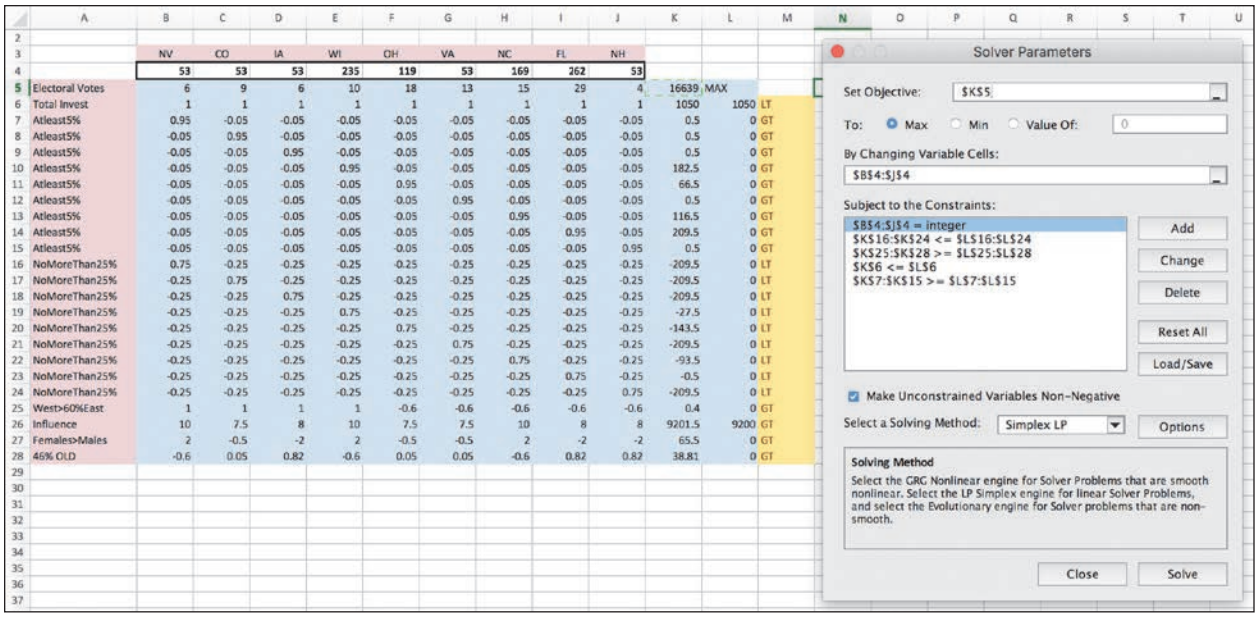
All told in algebraic terms, this integer programming model would have 9 decision variables and 24 constraints (one constraint for integer requirements).

Implementation

One approach would be to implement the model in strict "standard form," or a row-column form, where all constraints are written with decision variables on the left-hand side, a number on the right-hand side. Figure 8.8 shows such an implementation and displays the solved model.

Alternatively, we could use the spreadsheet to calculate different parts of the model in a less rigid manner, as well as uniquely implementing the repetitive constraints (b) and (c), and have a much more concise (but not as transparent) spreadsheet. This is shown in Figure 8.9.

LP models (and their specializations and generalizations) can also be specified directly in a number of other user-friendly modeling systems. Two of the best known are Lindo and Lingo (Lindo Systems, Inc., **lindo.com**; demos are available). Lindo is an



	A	B	C	D	E	F	G	H	I	J	K	L	M
3		NV	CO	IA	WI	OH	VA	NC	FL	NH			
4		53	53	53	235	119	53	169	262	53			
5	Electoral Votes	6	9	6	10	18	13	15	29	4	16639	MAX	
6	Total Invest	1	1	1	1	1	1	1	1	1	1050	1050	LT
7	Atleast5%	0.95	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	0.5	0	GT
8	Atleast5%	-0.05	0.95	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	0.5	0	GT
9	Atleast5%	-0.05	-0.05	0.95	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	0.5	0	GT
10	Atleast5%	-0.05	-0.05	-0.05	0.95	-0.05	-0.05	-0.05	-0.05	-0.05	182.5	0	GT
11	Atleast5%	-0.05	-0.05	-0.05	-0.05	0.95	-0.05	-0.05	-0.05	-0.05	66.5	0	GT
12	Atleast5%	-0.05	-0.05	-0.05	-0.05	-0.05	0.95	-0.05	-0.05	-0.05	0.5	0	GT
13	Atleast5%	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	0.95	-0.05	-0.05	116.5	0	GT
14	Atleast5%	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	0.95	-0.05	209.5	0	GT
15	Atleast5%	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	-0.05	0.95	0.5	0	GT
16	NoMoreThan25%	0.75	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-209.5	0	LT
17	NoMoreThan25%	-0.25	0.75	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-209.5	0	LT
18	NoMoreThan25%	-0.25	-0.25	0.75	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-209.5	0	LT
19	NoMoreThan25%	-0.25	-0.25	-0.25	0.75	-0.25	-0.25	-0.25	-0.25	-0.25	-27.5	0	LT
20	NoMoreThan25%	-0.25	-0.25	-0.25	-0.25	0.75	-0.25	-0.25	-0.25	-0.25	-143.5	0	LT
21	NoMoreThan25%	-0.25	-0.25	-0.25	-0.25	-0.25	0.75	-0.25	-0.25	-0.25	-209.5	0	LT
22	NoMoreThan25%	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	0.75	-0.25	-0.25	-93.5	0	LT
23	NoMoreThan25%	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	0.75	-0.25	-0.5	0	LT
24	NoMoreThan25%	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	0.75	-209.5	0	LT
25	West>60%East	1	1	1	1	-0.6	-0.6	-0.6	-0.6	-0.6	0.4	0	GT
26	Influence	10	7.5	8	10	7.5	7.5	10	8	8	9201.5	9200	GT
27	Females>Males	2	-0.5	-2	2	-0.5	-0.5	2	-2	-2	65.5	0	GT
28	46% OLD	-0.6	0.05	0.82	-0.6	0.05	0.05	-0.6	0.82	0.82	38.81	0	GT

FIGURE 8.8 **Model for Election Resource Allocation—Standard Version.**

LP and integer programming system. Models are specified in essentially the same way that they are defined algebraically. Based on the success of Lindo, the company developed Lingo, a modeling language that includes the powerful Lindo optimizer and extensions for solving nonlinear problems. Many other modeling languages such as AMPL, AIMMS, MPL, XPRESS, and others are available.

The most common optimization models can be solved by a variety of mathematical programming methods, including the following:

- Assignment (best matching of objects)
- Dynamic programming
- Goal programming
- Investment (maximizing rate of return)
- Linear and integer programming
- Network models for planning and scheduling
- Nonlinear programming
- Replacement (capital budgeting)
- Simple inventory models (e.g., economic order quantity)
- Transportation (minimize cost of shipments)

SECTION 8.6 REVIEW QUESTIONS

1. List and explain the assumptions involved in LP.
2. List and explain the characteristics of LP.
3. Describe an allocation problem.
4. Define the product-mix problem.
5. Define the blending problem.
6. List several common optimization models.

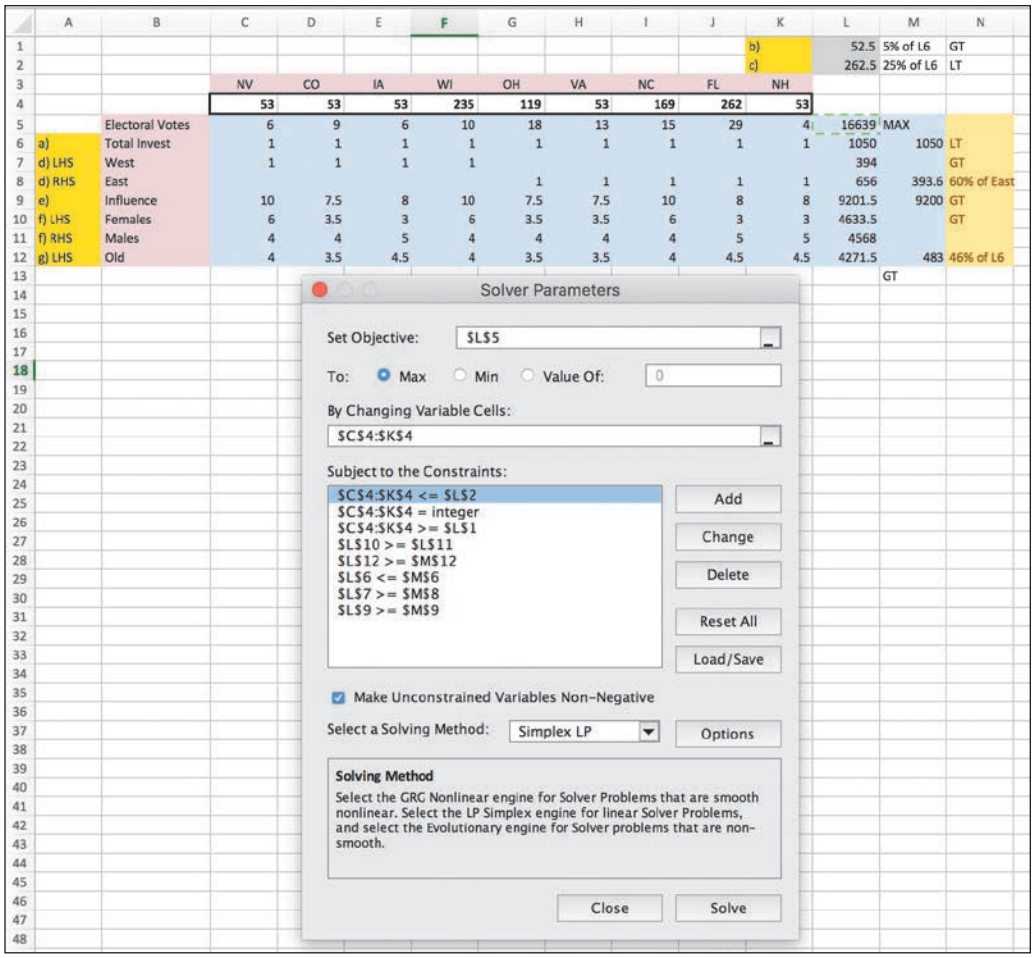


FIGURE 8.9 **A Compact Formulation for Election Resource Allocation.**

8.7 MULTIPLE GOALS, SENSITIVITY ANALYSIS, WHAT-IF ANALYSIS, AND GOAL SEEKING

Many, if not most, decision situations involve juggling between competing goals and alternatives. In addition, there is significant uncertainty about the assumptions and predictions being used in building a prescriptive analytics model. The following paragraphs simply recognize that these are also addressed in prescriptive analytics software and techniques. Coverage of these techniques is usually common in prescriptive analytics or operations research/management science courses.

Multiple Goals

The analysis of management decisions aims at evaluating, to the greatest possible extent, how far each alternative advances managers toward their goals. Unfortunately, managerial problems are seldom evaluated with a single simple goal, such as profit maximization. Today's management systems are much more complex, and one with a single goal is rare. Instead, managers want to attain *simultaneous goals*, some of which may conflict. Different stakeholders have different goals. Therefore, it is often necessary to analyze

each alternative in light of its determination of each of several goals (see Koksalan & Zionts, 2001).

For example, consider a profit-making firm. In addition to earning money, the company wants to grow, develop its products and employees, provide job security to its workers, and serve the community. Managers want to satisfy the shareholders and at the same time enjoy high salaries and expense accounts, and employees want to increase their take-home pay and benefits. When a decision is to be made—say, about an investment project—some of these goals complement each other, whereas others conflict. Kearns (2004) described how the analytic hierarchy process (AHP) combined with integer programming, addresses multiple goals in evaluating information technology (IT) investments.

Many quantitative models of decision theory are based on comparing a single measure of effectiveness, generally some form of utility to the decision maker. Therefore, it is usually necessary to transform a multiple-goal problem into a single-measure-of-effectiveness problem before comparing the effects of the solutions. This is a common method for handling multiple goals in an LP model.

Certain difficulties may arise when analyzing multiple goals:

- It is usually difficult to obtain an explicit statement of the organization's goals.
- The decision maker may change the importance assigned to specific goals over time or for different decision scenarios.
- Goals and subgoals are viewed differently at various levels of the organization and within different departments.
- Goals change in response to changes in the organization and its environment.
- The relationship between alternatives and their role in determining goals may be difficult to quantify.
- Complex problems are solved by groups of decision makers, each of whom has a personal agenda.
- Participants assess the importance (priorities) of the various goals differently.

Several methods of handling multiple goals can be used when working with such situations. The most common ones are

- Utility theory
- Goal programming
- Expression of goals as constraints, using LP
- A points system

Sensitivity Analysis

A model builder makes predictions and assumptions regarding input data, many of which deal with the assessment of uncertain futures. When the model is solved, the results depend on these data. **Sensitivity analysis** attempts to assess the impact of a change in the input data or parameters on the proposed solution (i.e., the result variable).

Sensitivity analysis is extremely important in prescriptive analytics because it allows flexibility and adaptation to changing conditions and to the requirements of different decision-making situations, provides a better understanding of the model and the decision-making situation it attempts to describe, and permits the manager to input data to increase the confidence in the model. Sensitivity analysis tests relationships such as the following:

- The impact of changes in external (uncontrollable) variables and parameters on the outcome variable(s)
- The impact of changes in decision variables on the outcome variable(s)

- The effect of uncertainty in estimating external variables
- The effects of different dependent interactions among variables
- The robustness of decisions under changing conditions

Sensitivity analyses are used for:

- Revising models to eliminate too-large sensitivities
- Adding details about sensitive variables or scenarios
- Obtaining better estimates of sensitive external variables
- Altering a real-world system to reduce actual sensitivities
- Accepting and using the sensitive (and hence vulnerable) real world, leading to the continuous and close monitoring of actual results

The two types of sensitivity analyses are automatic and trial and error.

AUTOMATIC SENSITIVITY ANALYSIS Automatic sensitivity analysis is performed in standard quantitative model implementations such as LP. For example, it reports the range within which a certain input variable or parameter value (e.g., unit cost) can vary without having any significant impact on the proposed solution. Automatic sensitivity analysis is usually limited to one change at a time, and only for certain variables. However, it is powerful because of its ability to establish ranges and limits very fast (and with little or no additional computational effort). Sensitivity analysis is provided by Solver and almost all other software packages such as Lindo. Consider the MBI Corporation example introduced previously. Sensitivity analysis could be used to determine that if the right-hand side of the marketing constraint on CC-8 could be decreased by one unit, then the net profit would increase by $1,333.33. This is valid for the right-hand side decreasing to zero. Significant additional analysis is possible along these lines.

TRIAL-AND-ERROR SENSITIVITY ANALYSIS The impact of changes in any variable, or in several variables, can be determined through a simple trial-and-error approach. You change some input data and solve the problem again. When the changes are repeated several times, better and better solutions may be discovered. Such experimentation, which is easy to conduct when using appropriate modeling software, such as Excel, has two approaches: what-if analysis and goal seeking.

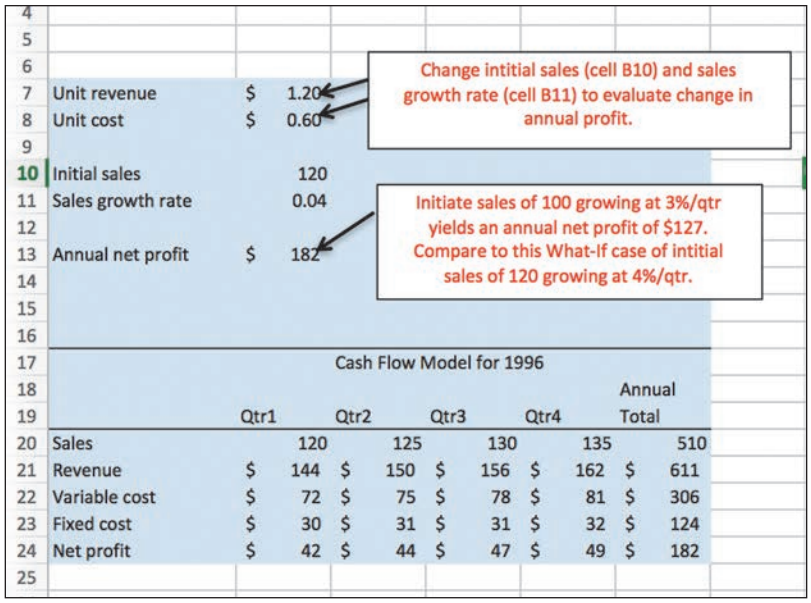
What-If Analysis

What-if analysis is structured as *What will happen to the solution if an input variable, an assumption, or a parameter value is changed?* Here are some examples:

- What will happen to the total inventory cost if the cost of carrying inventories increases by 10%?
- What will be the market share if the advertising budget increases by 5%?

With the appropriate user interface, it is easy for managers to ask a computer model these types of questions and get immediate answers. Furthermore, they can perform multiple cases and thereby change the percentage, or any other data in the question, as desired. The decision maker does all this directly, without a computer programmer.

Figure 8.10 shows a spreadsheet example of a what-if query for a cash flow problem. When the user changes the cells containing the initial sales (from 100 to 120) and the sales growth rate (from 3% to 4% per quarter), the program immediately recomputes the value of the annual net profit cell (from $127 to $182). At first, initial sales were 100, growing at 3% per quarter, yielding an annual net profit of $127. Changing the initial sales cell to 120 and the sales growth rate to 4% causes the annual net profit to rise to $182. What-if analysis is common in many decision systems. Users are given the opportunity to



Unit revenue	$ 1.20					
Unit cost	$ 0.60					
Initial sales	120					
Sales growth rate	0.04					
Annual net profit	$ 182					

Cash Flow Model for 1996					
	Qtr1	Qtr2	Qtr3	Qtr4	Annual Total
Sales	120	125	130	135	510
Revenue	$ 144	$ 150	$ 156	$ 162	$ 611
Variable cost	$ 72	$ 75	$ 78	$ 81	$ 306
Fixed cost	$ 30	$ 31	$ 31	$ 32	$ 124
Net profit	$ 42	$ 44	$ 47	$ 49	$ 182

FIGURE 8.10 **Example of a What-If Analysis Done in an Excel Worksheet.**

change their answers to some of the system's questions, and a revised recommendation is found.

Goal Seeking

Goal seeking calculates the values of the inputs necessary to achieve a desired level of an output (goal). It represents a backward solution approach. The following are some examples of goal seeking:

- What annual R&D budget is needed for an annual growth rate of 15% by 2018?
- How many nurses are needed to reduce the average waiting time of a patient in the emergency room to less than 10 minutes?

An example of goal seeking is shown in Figure 8.11. For example, in a financial planning model in Excel, the internal rate of return (IRR) is the interest rate that produces a net present value (NPV) of zero. Given a stream of annual returns in Column E, we can compute the NPV of planned investment. By applying goal seeking, we can determine the internal rate of return where the NPV is zero. The goal to be achieved is NPV equal to zero, which determines the internal rate of return of this cash flow, including the investment. We set the NPV cell to the value 0 by changing the interest rate cell. The answer is 38.77059%.

COMPUTING A BREAK-EVEN POINT BY USING GOAL SEEKING Some modeling software packages can directly compute break-even points, which is an important application of goal seeking. This involves determining the value of the decision variables (e.g., quantity to produce) that generate zero profit.

In many general applications programs, it can be difficult to conduct sensitivity analysis because the prewritten routines usually present only a limited opportunity for asking what-if questions. In a DSS, the what-if and the goal-seeking options must be easy to perform.

5						
6						
7	**Investment Problem**		Initial Investment:	$ 1,000.00		
8	**Example of GoalSeeking**		Interest Rate:	10%		
9						
10	**Find the Interest Rate**					
11	**(the Internal Rate of**	Year	Annual Returns		NPV Calculations	
12	**Return-IRR)**	1	$ 120.00		$109.09	
13	**that yields an NPV**	2	$ 130.00		$118.18	
14	**of $0**	3	$ 140.00		$127.27	
15		4	$ 150.00		$136.36	
16		5	$ 160.00		$145.45	
17		6	$ 152.00		$138.18	
18		7	$ 144.40		$131.27	
19		8	$ 137.18		$124.71	
20		9	$ 130.32		$118.47	
21		10	$ 123.80		$112.55	
22						
23			The NPV Solutions:		$261.55	
24						

FIGURE 8.11 Goal-Seeking Analysis.

SECTION 8.7 REVIEW QUESTIONS

1. List some difficulties that may arise when analyzing multiple goals.
2. List the reasons for performing sensitivity analysis.
3. Explain why a manager might perform what-if analysis.
4. Explain why a manager might use goal seeking.

8.8 DECISION ANALYSIS WITH DECISION TABLES AND DECISION TREES

Decision situations that involve a finite and usually not too large number of alternatives are modeled through an approach called **decision analysis** (see Arsham, 2006a,b; Decision Analysis Society, decision-analysis.society.informs.org). Using this approach, the alternatives are listed in a table or a graph, with their forecasted contributions to the goal(s) and the probability of obtaining the contribution. These can be evaluated to select the best alternative.

Single-goal situations can be modeled with *decision tables* or *decision trees*. Multiple goals (criteria) can be modeled with several other techniques, described later in this chapter.

Decision Tables

Decision tables conveniently organize information and knowledge in a systematic, tabular manner to prepare it for analysis. For example, say that an investment company is considering investing in one of three alternatives: bonds, stocks, or certificates of deposit (CDs). The company is interested in one goal: maximizing the yield on the investment after 1 year. If it were interested in other goals, such as safety or liquidity, the problem would be classified as one of *multicriteria decision analysis* (see Koksalan & Zionts, 2001).

The yield depends on the state of the economy sometime in the future (often called the *state of nature*), which can be in solid growth, stagnation, or inflation. Experts estimated the following annual yields:

- If there is solid growth in the economy, bonds will yield 12%, stocks 15%, and time deposits 6.5%.
- If stagnation prevails, bonds will yield 6%, stocks 3%, and time deposits 6.5%.
- If inflation prevails, bonds will yield 3%, stocks will bring a loss of 2%, and time deposits will yield 6.5%.

The problem is to select the one best investment alternative. These are assumed to be discrete alternatives. Combinations such as investing 50% in bonds and 50% in stocks must be treated as new alternatives.

The investment decision-making problem can be viewed as a *two-person game* (see Kelly, 2002). The investor makes a choice (i.e., a move), and then a state of nature occurs (i.e., makes a move). Table 8.3 shows the payoff of a mathematical model. The table includes *decision variables* (the alternatives), *uncontrollable variables* (the states of the economy; e.g., the environment), and *result variables* (the projected yield; e.g., outcomes). All the models in this section are structured in a spreadsheet framework.

If this were a decision-making problem under certainty, we would know what the economy would be and could easily choose the best investment. But that is not the case, so we must consider the two situations of uncertainty and risk. For uncertainty, we do not know the probabilities of each state of nature. For risk, we assume that we know the probabilities with which each state of nature will occur.

TREATING UNCERTAINTY Several methods are available for handling uncertainty. For example, the *optimistic approach* assumes that the best possible outcome of each alternative will occur and then selects the best of the best (i.e., stocks). The *pessimistic approach* assumes that the worst possible outcome for each alternative will occur and selects the best of these (i.e., CDs). Another approach simply assumes that all states of nature are equally possible (see Clemen & Reilly, 2000; Goodwin & Wright, 2000; Kontoghiorghes, Rustem, & Siokos, 2002). Every approach for handling uncertainty has serious problems. Whenever possible, the analyst should attempt to gather enough information so that the problem can be treated under assumed certainty or risk.

TREATING RISK The most common method for solving this risk analysis problem is to select the alternative with the greatest expected value. Assume that experts estimate the chance of solid growth at 50%, the chance of stagnation at 30%, and the chance of inflation at 20%. The decision table is then rewritten with the known probabilities (see Table 8.3). An expected value is computed by multiplying the results (i.e., outcomes) by their respective probabilities and adding them. For example, investing in bonds yields an expected return of $12(0.5) + 6(0.3) + 3(0.2) = 8.4\%$.

This approach can sometimes be a dangerous strategy because the utility of each potential outcome may be different from the value. Even if there is an infinitesimal chance of a catastrophic loss, the expected value may seem reasonable, but the investor may not be willing to cover the loss. For example, suppose a financial advisor presents you with

TABLE 8.3 Investment Problem Decision Table Model

	State of Nature (Uncontrollable Variables)		
Alternative	**Solid Growth (%)**	**Stagnation (%)**	**Inflation (%)**
Bonds	12.0	6.0	3.0
Stocks	15.0	3.0	–2.0
CDs	6.5	6.5	6.5

an "almost sure" investment of $1,000 that can double your money in one day, and then the advisor says, "Well, there is a 0.9999 probability that you will double your money, but unfortunately there is a 0.0001 probability that you will be liable for a $500,000 out-of-pocket loss." The expected value of this investment is as follows:

$$0.9999\ (\$2{,}000 - \$1{,}000) + .0001(-\$500{,}000 - \$1{,}000) = \$999.90 - \$50.10$$
$$= \$949.80$$

The potential loss could be catastrophic for any investor who is not a billionaire. Depending on the investor's ability to cover the loss, an investment has different expected utilities. Remember that the investor makes the decision only *once*.

Decision Trees

An alternative representation of the decision table is a decision tree. A **decision tree** shows the relationships of the problem graphically and can handle complex situations in a compact form. However, a decision tree can be cumbersome if there are many alternatives or states of nature. TreeAge Pro (TreeAge Software Inc., **treeage.com**) and PrecisionTree (Palisade Corp., **palisade.com**) include powerful, intuitive, and sophisticated decision tree analysis systems. These vendors also provide excellent examples of decision trees used in practice. Note that the phrase *decision tree* has been used to describe two different types of models and algorithms. In the current context, decision trees refer to scenario analysis. On the other hand, some classification algorithms in predictive analysis (see Chapters 4 and 5) are also called decision tree algorithms. The reader is advised to note the difference between two different uses of the same name – decision tree.

A simplified investment case of **multiple goals** (a decision situation in which alternatives are evaluated with several, sometimes conflicting, goals) is shown in Table 8.4. The three goals (criteria) are yield, safety, and liquidity. This situation is under assumed certainty; that is, only one possible consequence is projected for each alternative; the more complex cases of risk or uncertainty could be considered. Some of the results are qualitative (e.g., low, high) rather than numeric.

See Clemen and Reilly (2000), Goodwin and Wright (2000), and Decision Analysis Society (**informs.org/Community/DAS**) for more on decision analysis. Although doing so is quite complex, it is possible to apply mathematical programming directly to decision-making situations under risk. We discuss several other methods of treating risk later in the book. These include simulation, certainty factors, and fuzzy logic.

SECTION 8.8 REVIEW QUESTIONS

1. What is a decision table?
2. What is a decision tree?
3. How can a decision tree be used in decision making?
4. Describe what it means to have multiple goals.

TABLE 8.4 Multiple Goals

Alternative	Yield (%)	Safety	Liquidity
Bonds	8.4	High	High
Stocks	8.0	Low	High
CDs	6.5	Very high	High

8.9 INTRODUCTION TO SIMULATION

In this section and the next we introduce a category of techniques that are used for supporting decision making. Very broadly, these methods fall under the umbrella of simulation. **Simulation** is the appearance of reality. In decision systems, simulation is a technique for conducting experiments (e.g., what-if analyses) with a computer on a model of a management system. Strictly speaking, simulation is a *descriptive* rather than a *prescriptive* method. There is no automatic search for an optimal solution. Instead, a simulation model describes or predicts the characteristics of a given system under different conditions. When the values of the characteristics are computed, the best of several alternatives can be selected. The simulation process usually repeats an experiment many times to obtain an estimate (and a variance) of the overall effect of certain actions. For most situations, a computer simulation is appropriate, but there are some well-known manual simulations (e.g., a city police department simulated its patrol car scheduling with a carnival game wheel).

Typically, real decision-making situations involve some randomness. Because many decision situations deal with semistructured or unstructured situations, reality is complex, which may not be easily represented by optimization or other models but can often be handled by simulation. Simulation is one of the most commonly used decision support methods. See Application Case 8.6 for an example. Application Case 8.7 illustrates the value of simulation in another setting where the problem complexity does not permit building a traditional optimization model.

Major Characteristics of Simulation

Simulation typically involves building a model of reality to the extent practical. Simulation models may suffer from fewer assumptions about the decision situation as compared to other prescriptive analytic models. In addition, simulation is a technique for *conducting experiments*. Therefore, it involves testing specific values of the decision or uncontrollable variables in the model and observing the impact on the output variables.

Finally, simulation is normally used only when a problem is too complex to be treated using numerical optimization techniques. Complexity in this situation means either that the problem cannot be formulated for optimization (e.g., because the assumptions do not hold), that the formulation is too large, that there are too many interactions among the variables, or that the problem is stochastic in nature (i.e., exhibits risk or uncertainty).

Application Case 8.7 Steel Tubing Manufacturer Uses a Simulation-Based Production Scheduling System

A steel manufacturing plant produces rolled-steel tubes for different industries across the country. They build tubes based on a customer's requirements and specifications. Maintaining high-quality norms and timely delivery of products are two of the foremost important criteria for this steel tubing plant. The plant views its manufacturing system as a sequence of operations where it unrolls steel from one reel and rolls it onto a different reel. This happens once the forming, welding, editing, or inspecting operation is finished. The ultimate product would be a reel of rolled steel tubing that weighs about 20 tons. The reel is then shipped to the customer.

A key challenge for management is to be able to predict the appropriate delivery date for an order, and its impact on the currently planned production schedule. Given the complexity of the production process, it is not easy to develop an optimization model in Excel or other software to build a

(*Continued*)

Application Case 8.7 (Continued)

production schedule (see Application Case 8.1). The issue is that these tools fail to capture key planning issues such as employee schedules and qualifications, material accessibility, material allocation complication, and random aspects of the operation.

Methodology/Solution

When traditional modeling methods do not capture the problem subtleties or complexities, a simulation model could perhaps be built. The predictive analysis approach uses a versatile Simio simulation model that takes into consideration all the operational complexity, manufacturing material matching algorithms, and deadline considerations. Also, Simio's service offering, known as risk-based planning and scheduling (RPS), provides some user interfaces and reports simply designed for production management. This gives the client the ability to explore the impact of a new order on their production plan and schedule within about 10 minutes.

Results/Benefits

Such models provide significant visibility into the production schedule. The risk-based planning and scheduling system should be able to warn the master scheduler that a specific order has a chance of being delivered late. Changes could also be made sooner to rectify issues with an order. Success for this steel tubing manufacturer is directly tied to product quality and on-time delivery. By exploitation of Simio's predictive RPS offering, the plant expects improved market share.

Questions for Discussion

1. Explain the advantages of using Simio's simulation model over traditional methods.
2. In what ways has the predictive analysis approach helped management achieve the goals of analyzing the production schedules?
3. Besides the steel manufacturing industry, in what other industries could such a modeling approach help improve quality and service?

What Can We Learn from This Application Case?

By using Simio's simulation model, the manufacturing plant made better decisions in assessment of operations, taking all of the problem issues into consideration. Thus, a simulation-based production scheduling system could derive higher returns and market share for the steel tubing manufacturer. Simulation is an important technique for prescriptive analytics.

Compiled from Arthur, Molly. "Simulation-Based Production Scheduling System." **www.simio.com**, Simio LLC, 2014, **www.simio.com/case-studies/A-Steel-Tubing-Manufacturer-Expects-More-Market-Share/A-Steel-Tubing-Manufacturer-Expects-More-Market-Share.pdf** (accessed September 2018); "Risk-Based Planning and Scheduling (RPS) with Simio." **www.simio.com**, Simio LLC, **www.simio.com/about-simio/why-simio/simio-RPS-risk-based-planning-and-scheduling.php** (accessed September 2018).

Advantages of Simulation

Simulation is used in decision support modeling for the following reasons:

- The theory is fairly straightforward.
- A great amount of *time compression* can be attained, quickly giving a manager some feel as to the long-term (1- to 10- year) effects of many policies.
- Simulation is descriptive rather than normative. This allows the manager to pose what-if questions. Managers can use a trial-and-error approach to problem solving and can do so faster, at less expense, more accurately, and with less risk.
- A manager can experiment to determine which decision variables and which parts of the environment are really important, and with different alternatives.
- An accurate simulation model requires an intimate knowledge of the problem, thus forcing the model builder to constantly interact with the manager. This is desirable

for DSS development because the developer and manager both gain a better understanding of the problem and the potential decisions available.
- The model is built from the manager's perspective.
- The simulation model is built for one particular problem and typically cannot solve any other problem. Thus, no generalized understanding is required of the manager; every component in the model corresponds to part of the real system.
- Simulation can handle an extremely wide variety of problem types, such as inventory and staffing, as well as higher-level managerial functions, such as long-range planning.
- Simulation generally can include the real complexities of problems; simplifications are not necessary. For example, simulation can use real probability distributions rather than approximate theoretical distributions.
- Simulation automatically produces many important performance measures.
- Simulation is often the only DSS modeling method that can readily handle relatively unstructured problems.
- Some relatively easy-to-use simulation packages (e.g., Monte Carlo simulation) are available. These include add-in spreadsheet packages (e.g., @RISK), influence diagram software, Java-based (and other Web development) packages, and the visual interactive simulation systems to be discussed shortly.

Disadvantages of Simulation

The primary disadvantages of simulation are as follows:

- An optimal solution cannot be guaranteed, but relatively good ones are generally found.
- Simulation model construction can be a slow and costly process, although newer modeling systems are easier to use than ever.
- Solutions and inferences from a simulation study are usually not transferable to other problems because the model incorporates unique problem factors.
- Simulation is sometimes so easy to explain to managers that analytic methods are often overlooked.
- Simulation software sometimes requires special skills because of the complexity of the formal solution method.

The Methodology of Simulation

Simulation involves setting up a model of a real system and conducting repetitive experiments on it. The methodology consists of the following steps, as shown in Figure 8.12:

1. ***Define the problem.*** We examine and classify the real-world problem, specifying why a simulation approach is appropriate. The system's boundaries, environment, and other such aspects of problem clarification are handled here.
2. ***Construct the simulation model.*** This step involves determination of the variables and their relationships, as well as data gathering. Often the process is described by using a flowchart, and then a computer program is written.
3. ***Test and validate the model.*** The simulation model must properly represent the system being studied. Testing and validation ensure this.
4. ***Design the experiment.*** When the model has been proven valid, an experiment is designed. Determining how long to run the simulation is part of this step. There are two important and conflicting objectives: accuracy and cost. It is also prudent to identify typical (e.g., mean and median cases for random variables), best-case (e.g., low-cost, high-revenue), and worst-case (e.g., high-cost, low-revenue) scenarios.

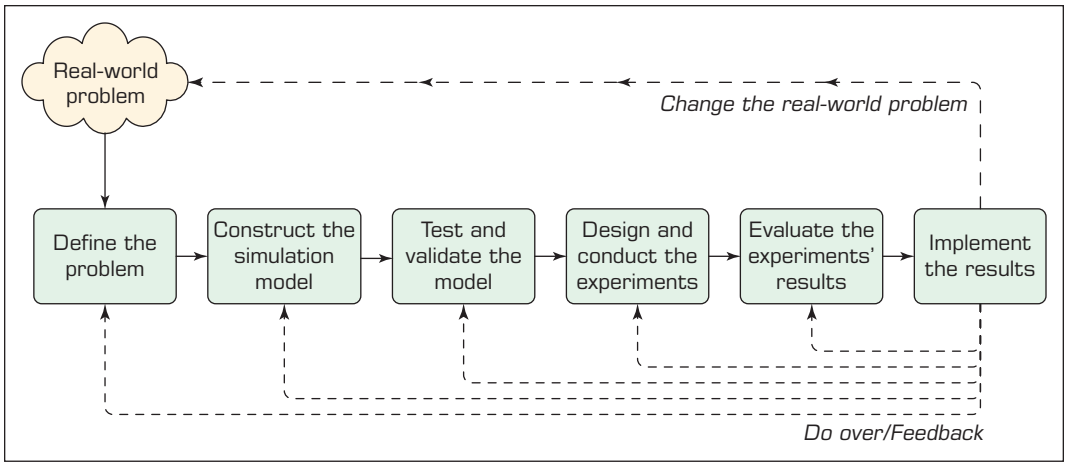


FIGURE 8.12 **The Process of Simulation.**

These help establish the ranges of the decision variables and environment in which to work and also assist in debugging the simulation model.

5. ***Conduct the experiment.*** Conducting the experiment involves issues ranging from random-number generation to result presentation.
6. ***Evaluate the results.*** The results must be interpreted. In addition to standard statistical tools, sensitivity analyses can also be used.
7. ***Implement the results.*** The implementation of simulation results involves the same issues as any other implementation. However, the chances of success are better because the manager is usually more involved with the simulation process than with other models. Higher levels of managerial involvement generally lead to higher levels of implementation success.

Banks and Gibson (2009) presented some useful advice about simulation practices. For example, they list the following seven issues as the common mistakes committed by simulation modelers. The list, though not exhaustive, provides general directions for professionals working on simulation projects.

- Focusing more on the model than on the problem
- Providing point estimates
- Not knowing when to stop
- Reporting what the client wants to hear rather than what the model results say
- Lack of understanding of statistics
- Confusing cause and effect
- Failure to replicate reality

In a follow-up article they provide additional guidelines. You should consult this article: **analytics-magazine.org/spring-2009/205-software-solutions-the-abcs-of-simulation-practice.html**.

Simulation Types

As we have seen, simulation and modeling are used when pilot studies and experimenting with real systems are expensive or sometimes impossible. Simulation models allow us to investigate various interesting scenarios before making any investment. In fact, in simulations, the real-world operations are mapped into the simulation model. The model consists of relationships and, consequently, equations that all together present the

real-world operations. The results of a simulation model, then, depend on the set of parameters given to the model as inputs.

There are various simulation paradigms such as Monte Carlo simulation, discrete event, agent based, or system dynamics. One of the factors that determine the type of simulation technique is the level of abstraction in the problem. Discrete events and agent-based models are usually used for middle or low levels of abstraction. They usually consider individual elements such as people, parts, and products in the simulation models, whereas systems dynamics is more appropriate for aggregate analysis.

In the following section, we introduce the major types of simulation: probabilistic simulation, time-dependent and time-independent simulation, and visual simulation. There are many other simulation techniques such as system dynamics modeling, and agent-based modeling. As has been noted before, the goal here is to make you aware of the potential of some of these techniques as opposed to make you an expert in using them.

PROBABILISTIC SIMULATION In probabilistic simulation, one or more of the independent variables (e.g., the demand in an inventory problem) are probabilistic. They follow certain probability distributions, which can be either discrete distributions or continuous distributions:

- *Discrete distributions* involve a situation with a limited number of events (or variables) that can take on only a finite number of values.
- *Continuous distributions* are situations with unlimited numbers of possible events that follow density functions, such as the normal distribution.

The two types of distributions are shown in Table 8.5.

TIME-DEPENDENT VERSUS TIME-INDEPENDENT SIMULATION *Time-independent* refers to a situation in which it is not important to know exactly when the event occurred. For example, we may know that the demand for a certain product is three units per day, but we do not care *when* during the day the item is demanded. In some situations, time may not be a factor in the simulation at all, such as in steady-state plant control design. However, in waiting-line problems applicable to e-commerce, it is important to know the precise time of arrival (to know whether the customer will have to wait). This is a *time-dependent* situation.

Monte Carlo Simulation

In most business decision problems, we usually employ one of the following two types of probabilistic simulations. The most common simulation method for business decision problems is the **Monte Carlo simulation**. This method usually begins with building a

TABLE 8.5 Discrete versus Continuous Probability Distributions

Daily Demand	Discrete Probability	Continuous Probability
5	0.10	Daily demand is normally distributed with a mean of 7 and a standard deviation of 1.2
6	0.15	
7	0.30	
8	0.25	
9	0.20	

model of the decision problem without having to consider the uncertainty of any variables. Then we recognize that certain parameters or variables are uncertain or follow an assumed or estimated probability distribution. This estimation is based on analysis of past data. Then we begin running sampling experiments. Running sampling experiments consists of generating random values of uncertain parameters and then computing values of the variables that are impacted by such parameters or variables. These sampling experiments essentially amount to solving the same model hundreds or thousands of times. We can then analyze the behavior of these dependent or performance variables by examining their statistical distributions. This method has been used in simulations of physical as well as business systems. A good public tutorial on the Monte Carlo simulation method is available on **Palisade.com** (**http://www.palisade.com/risk/monte_carlo_simulation.asp**). Palisade markets a tool called @RISK, a popular spreadsheet-based Monte Carlo simulation software. Another popular software in this category is Crystal Ball, now marketed by Oracle as Oracle Crystal Ball. Of course, it is also possible to build and run Monte Carlo experiments within an Excel spreadsheet without using any add-on software such as the two just mentioned. But these tools make it more convenient to run such experiments in Excel-based models. Monte Carlo simulation models have been used in many commercial applications. Examples include Procter & Gamble using these models to determine hedging foreign-exchange risks; Lilly using the model for deciding optimal plant capacity; Abu Dhabi Water and Electricity Company using @Risk for forecasting water demand in Abu Dhabi; and literally thousands of other actual case studies. Each of the simulation software companies' Web sites include many such success stories.

Discrete Event Simulation

Discrete event simulation refers to building a model of a system where the interaction between different entities is studied. The simplest example of this is a shop consisting of a server and customers. By modeling the customers arriving at various rates and the server serving at various rates, we can estimate the average performance of the system, waiting time, the number of waiting customers, and so on. Such systems are viewed as collections of customers, queues, and servers. There are thousands of documented applications of discrete event simulation models in engineering, business, and so on. Tools for building discrete event simulation models have been around for a long time, but these have evolved to take advantage of developments in graphical capabilities for building and understanding the results of such simulation models. We will discuss this modeling method further in the next section. Application Case 8.8 gives an example of the use of such simulation in analyzing complexities of a supply chain that uses a visual simulation to be described in the next section.

Application Case 8.8 Cosan Improves Its Renewable Energy Supply Chain Using Simulation

Introduction

Cosan is a Brazil-based conglomerate that operates globally. One of its major activities is to grow and process sugar cane. Besides being a major source of sugar, sugar cane is now a major source of ethanol, a main ingredient in renewable energy. Because of the growing demand for renewable energy, ethanol production has become such a major activity for Cosan that it now operates two refineries in addition to 18 production plants, and of course, millions of hectares of sugar cane farms. According to recent data, it processed over 44 million tons of sugar cane, produced over 1.3 billion liters of ethanol, and produced 3.3 million tons of sugar. As one might imagine, operations of this scale lead to complex supply chains. So the logistics team was asked

to make recommendations to the senior management to:

- Determine the optimum number of vehicles required in a fleet used to transport sugar cane to processing mills to preserve capital.
- Propose how to increase the actual capacity of sugar cane received at the sugar mills.
- Identify the production bottleneck problems to solve to improve the flow of sugar cane.

Methodology/Solution

The logistics team worked with Simio software and built a complex simulation model of the Cosan supply chain as it pertains to these issues. According to a Simio brief, "Over the course of three months, newly hired engineers collected data in the field and received hands-on training and modeling assistance from Paragon Consulting of San Palo."

To model agricultural operations to analyze the sugar cane's postharvest journey to production mills, the model objectives included details of the fleet of road transport sugar cane crop to Unity Costa Pinto, the actual capacity of reception of cane sugar mills, bottlenecks and points for improvement in the flow of CCT (cut-load-haul) of cane sugar, and so on.

The model parameters are as follows:

Input Variables: 32
Output Variables: 39
Auxiliary Variables: 92
Variable Entities: 8
Input Tables: 19
Simulated Days: 240 (1st season)
Number of Entities: 12 (10 harvester compositional types for transport of sugar cane)

Results/Benefits

Analyses produced by these Simio models provided a good view of the risk of operation over the 240-day period due to various uncertainties. By analyzing the various bottlenecks and ways to mitigate those scenarios, the company was able to make better decisions and save over $500,000 from this modeling effort alone.

Questions for Discussion

1. What type of supply chain disruptions might occur in moving the sugar cane from the field to the production plants to develop sugar and ethanol?
2. What types of advanced planning and prediction might be useful in mitigating such disruptions?

What Can We Learn from This Application Case?

This short application story illustrates the value of applying simulation to a problem where it might be difficult to build an optimization model. By incorporating a discrete event simulation model and visual interactive simulation (VIS), one can visualize the impact of interruptions in supply chain due to fleet failure, unexpected downtime at the plant, and so on, and come up with planned corrections.

Sources: Compiled from Wikipedia contributors, Cosan, *Wikipedia, The Free Encyclopedia,* **https://en.wikipedia.org/w/index.php?title=Cosan&oldid=713298536** (accessed July 10, 2016); Agricultural Operations Simulation Case Study: Cosan, **http://www.simio.com/case-studies/Cosan-agricultural-logistics-simulation-software-case-study/agricultural-simulation-software-case-study-video-cosan.php** (accessed July 2016); Cosan Case Study: Optimizing agricultural logistics operations, **http://www.simio.com/case-studies/Cosan-agricultural-logistics-simulation-software-case-study/index.php** (accessed July 2016).

SECTION 8.9 REVIEW QUESTIONS

1. List the characteristics of simulation.
2. List the advantages and disadvantages of simulation.
3. List and describe the steps in the methodology of simulation.
4. List and describe the types of simulation.

8.10 VISUAL INTERACTIVE SIMULATION

We next examine methods that show a decision maker a representation of the decision-making situation in action as it runs through scenarios of the various alternatives. These powerful methods overcome some of the inadequacies of conventional methods and help build trust in the solution attained because they can be visualized directly.

Conventional Simulation Inadequacies

Simulation is a well-established, useful, descriptive, mathematics-based method for gaining insight into complex decision-making situations. However, simulation does not usually allow decision makers to see how a solution to a complex problem evolves over (compressed) time, nor can decision makers interact with the simulation (which would be useful for training purposes and teaching). Simulation generally reports statistical results at the end of a set of experiments. Decision makers are thus not an integral part of simulation development and experimentation, and their experience and judgment cannot be used directly. If the simulation results do not match the intuition or judgment of the decision maker, a *confidence gap* in the results can occur.

Visual Interactive Simulation

Visual interactive simulation (VIS), also known as **visual interactive modeling (VIM)** and *visual interactive problem solving*, is a simulation method that lets decision makers see what the model is doing and how it interacts with the decisions made, as they are made. This technique has been used with great success in operations analysis in many fields such as supply chain and healthcare. The user can employ his or her knowledge to determine and try different decision strategies while interacting with the model. Enhanced learning, about both the problem and the impact of the alternatives tested, can and does occur. Decision makers also contribute to model validation. Decision makers who use VIS generally support and trust their results.

VIS uses animated computer graphic displays to present the impact of different managerial decisions. It differs from regular graphics in that the user can adjust the decision-making process and see results of the intervention. A visual model is a graphic used as an integral part of decision making or problem solving, not just as a communication device. Some people respond better than others to graphical displays, and this type of interaction can help managers learn about the decision-making situation.

VIS can represent static or dynamic systems. Static models display a visual image of the result of one decision alternative at a time. Dynamic models display systems that evolve over time, and the evolution is represented by animation. The latest visual simulation technology has been coupled with the concept of virtual reality, where an artificial world is created for a number of purposes, from training to entertainment to viewing data in an artificial landscape. For example, the U.S. military uses VIS systems so that ground troops can gain familiarity with terrain or a city to very quickly orient themselves. Pilots also use VIS to gain familiarity with targets by simulating attack runs. The VIS software can also include GIS coordinates.

Visual Interactive Models and DSS

VIM in DSS has been used in several operations management decisions. The method consists of priming (like priming a water pump) a visual interactive model of a plant (or company) with its current status. The model then runs rapidly on a computer, allowing managers to observe how a plant is likely to operate in the future.

Waiting-line management (queuing) is a good example of VIM. Such a DSS usually computes several measures of performance for the various decision alternatives

(e.g., waiting time in the system). Complex waiting-line problems require simulation. VIM can display the size of the waiting line as it changes during the simulation runs and can also graphically present the answers to what-if questions regarding changes in input variables. Application Case 8.9 gives an example of a visual simulation that was used to explore the applications of radio-frequency identification (RFID) technology in developing new scheduling rules in a manufacturing setting.

The VIM approach can also be used in conjunction with artificial intelligence. Integration of the two techniques adds several capabilities that range from the ability to build systems graphically to learning about the dynamics of the system. These systems, especially those developed for the military and the video-game industry, have "thinking" characters who can behave with a relatively high level of intelligence in their interactions with users.

Simulation Software

Hundreds of simulation packages are available for a variety of decision-making situations. Many run as Web-based systems. *ORMS Today* publishes a periodic review of simulation software. One recent review (current as of October 2018) is located at **https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-44-Number-5/Simulation-Software-Survey-Simulation-new-and-improved-reality-show** (accessed November 2018). PC software packages include Analytica (Lumina Decision Systems, **lumina.com**) and the Excel add-ins Crystal Ball (now sold by Oracle as Oracle Crystal Ball, **oracle.com**) and @RISK (Palisade Corp., **palisade.com**). A major commercial software for discrete event simulation has been Arena (sold by Rockwell Intl., **arenasimulation.com**). Original developers of Arena have now developed Simio (**simio.com**), a user-friendly VIS software. Another popular discrete event VIS software is ExtendSim (**extendsim.com**). SAS has a graphical analytics software package called JMP that also includes a simulation component in it.

Application Case 8.9

Improving Job-Shop Scheduling Decisions through RFID: A Simulation-Based Assessment

A manufacturing services provider of complex optical and electromechanical components seeks to gain efficiency in its job-shop scheduling decision because the current shop-floor operations suffer from a few issues:

- There is no system to record when the work-in-process (WIP) items actually arrive at or leave operating workstations and how long those WIPs actually stay at each workstation.
- The current system cannot monitor or keep track of the movement of each WIP in the production line in real time.

As a result, the company is facing two main issues at this production line: high backlogs and high costs of overtime to meet the demand. In addition, the upstream cannot respond to unexpected incidents such as changes in demand or material shortages quickly enough and revise schedules in a cost-effective manner. The company is considering implementing RFID on a production line. However, the company does not know if going to this major expense of adding RFID chips on production boxes, installing RFID readers throughout the production line, and of course, the systems to process this information will result in any real gains. So one question is to explore any new production scheduling changes that may result by investing in RFID infrastructure.

Methodology

Because exploring the introduction of any new system in the physical production system can be extremely expensive or even disruptive, a discrete event simulation model was developed to examine how tracking and traceability through RFID can facilitate job-shop production scheduling activities. A visibility-based scheduling (VBS) rule that utilizes the real-time traceability systems to track those WIPs, parts and components, and raw

(*Continued*)

Application Case 8.9 (Continued)

materials in shop-floor operations was proposed. A simulation approach was applied to examine the benefit of the VBS rule against the classical scheduling rules: the first-in-first-out and earliest due date dispatching rules. The simulation model was developed using Simio. Simio is a 3-D simulation modeling software package that employs an object-oriented approach to modeling and has recently been used in many areas such as factories, supply chains, healthcare, airports, and service systems.

Figure 8.13 presents a screenshot of the Simio interface panel of this production line. The parameter estimates used for the initial state in the simulation model include weekly demand and forecast, process flow, number of workstations, number of shop-floor operators, and operating time at each workstation. In addition, parameters of some of the input data such as RFID tagging time, information retrieving time, or system updating time are estimated from a pilot study and from the subject matter experts. Figure 8.14 presents the process view of the simulation model where specific simulation commands are implemented and coded. Figures 8.15 and 8.16 present the standard report view and pivot grid report of the simulation model. The standard report and pivot grid format provide a very quick method to find specific statistical results such as average, percent, total, maximum, or minimum values of variables assigned and captured as an output of the simulation model.

Results

The results of the simulation suggest that an RFID-based scheduling rule generates better performance compared to traditional scheduling rules with regard to processing time, production time, resource utilization, backlogs, and productivity. The company can take these productivity gains and perform cost/benefit analyses in making the final investment decisions.

FIGURE 8.13 Simio Interface View of the Simulation System. *Source:* Used with permission from Simio LLC.

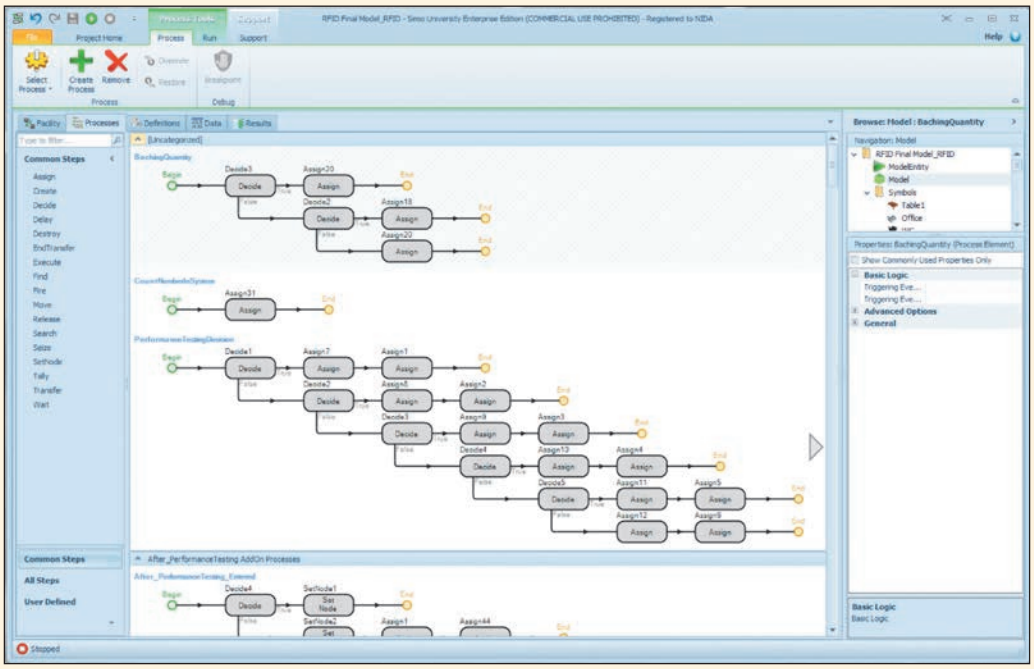


FIGURE 8.14 Process View of the Simulation Model. *Source:* Used with permission from Simio LLC.

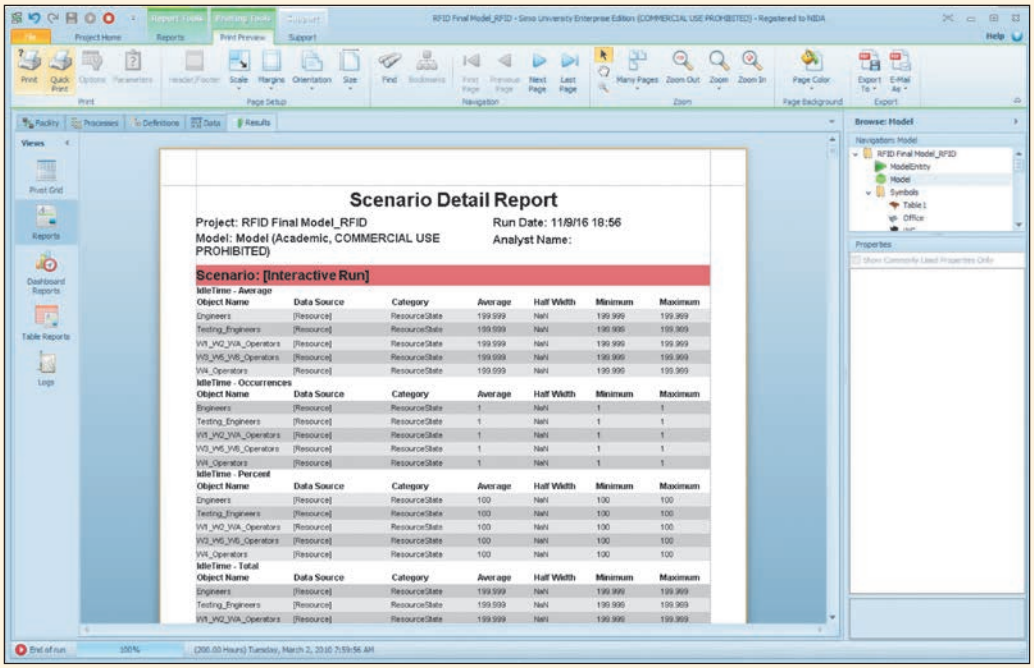


FIGURE 8.15 Standard Report View. *Source:* Used with permission from Simio LLC.

(*Continued*)

Application Case 8.9 (Continued)

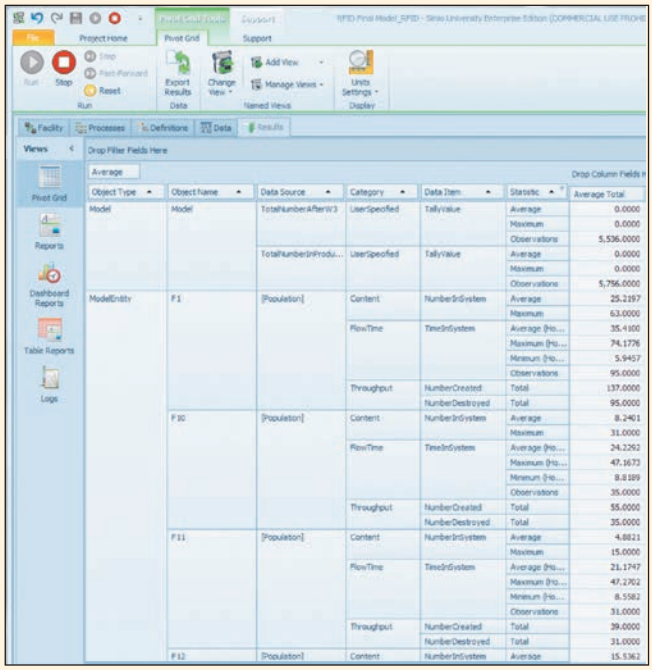


FIGURE 8.16 Pivot Grid Report from a Simio Run. *Source:* Used with permission from Simio LLC.

Questions for Discussion

1. In situations such as what this case depicts, what other approaches can one take to analyze investment decisions?
2. How would one save time if an RFID chip can tell the exact location of a product in process?
3. Research to learn about the applications of RFID sensors in other settings. Which one do you find most interesting?

Source: Based on Chongwatpol, J., & Sharda, R. (2013). RFID-enabled track and traceability in job-shop scheduling environment. *European Journal of Operational Research, 227*(3), 453–463, **http://dx.doi.org/10.1016/j.ejor.2013.01.009**.

For information about simulation software, see the Society for Modeling and Simulation International (scs.org) and the annual software surveys at *ORMS Today* (**https://www.informs.org/ORMS-Today/**).

SECTION 8.10 REVIEW QUESTIONS

1. Define *visual simulation* and compare it to conventional simulation.
2. Describe the features of VIS (i.e., VIM) that make it attractive for decision makers.
3. How can VIS be used in operations management?
4. How is an animated film like a VIS application?

Chapter Highlights

- Models play a major role in DSS because they are used to describe real decision-making situations. There are several types of models.
- Models can be static (i.e., a single snapshot of a situation) or dynamic (i.e., multiperiod).
- Analysis is conducted under assumed certainty (which is most desirable), risk, or uncertainty (which is least desirable).
- Influence diagrams graphically show the interrelationships of a model. They can be used to enhance the use of spreadsheet technology.
- Spreadsheets have many capabilities, including what-if analysis, goal seeking, programming, database management, optimization, and simulation.
- Decision tables and decision trees can model and solve simple decision-making problems.
- Mathematical programming is an important optimization method.
- LP is the most common mathematical programming method. It attempts to find an optimal allocation of limited resources under organizational constraints.
- The major parts of an LP model are the objective function, the decision variables, and the constraints.
- Multicriteria decision-making problems are difficult but not impossible to solve.
- What-if and goal seeking are the two most common methods of sensitivity analysis.
- Many DSS development tools include built-in quantitative models (e.g., financial, statistical) or can easily interface with such models.
- Simulation is a widely used DSS approach that involves experimentation with a model that represents the real decision-making situation.
- Simulation can deal with more complex situations than optimization, but it does not guarantee an optimal solution.
- There are many different simulation methods. Some that are important for decision making include Monte Carlo simulation and discrete event simulation.
- VIS/VIM allows a decision maker to interact directly with a model and shows results in an easily understood manner.

Key Terms

certainty
decision analysis
decision table
decision tree
decision variable
discrete event simulation
dynamic models
environmental scanning and analysis
forecasting
goal seeking
influence diagram
intermediate result variable
linear programming (LP)
mathematical programming
Monte Carlo simulation
multidimensional analysis (modeling)
multiple goals
optimal solution
parameter
quantitative model
result (outcome) variable
risk
risk analysis
sensitivity analysis
simulation
static models
uncertainty
uncontrollable variable
visual interactive modeling (VIM)
visual interactive simulation (VIS)
what-if analysis

Questions for Discussion

1. How does prescriptive analytics relate to descriptive and predictive analytics?
2. Explain the differences between static and dynamic models. How can one evolve into the other?
3. What is the difference between an optimistic approach and a pessimistic approach to decision making under assumed uncertainty?
4. Explain why solving problems under uncertainty sometimes involves assuming that the problem is to be solved under conditions of risk.
5. Excel is probably the most popular spreadsheet software for PCs. Why? What can we do with this package that makes it so attractive for modeling efforts?

6. Explain how decision trees work. How can a complex problem be solved by using a decision tree?
7. Explain how LP can solve allocation problems.
8. What are the advantages of using a spreadsheet package to create and solve LP models? What are the disadvantages?
9. What are the advantages of using an LP package to create and solve LP models? What are the disadvantages?
10. What is the difference between decision analysis with a single goal and decision analysis with multiple goals (i.e., criteria)? Explain the difficulties that may arise when analyzing multiple goals.
11. Explain how multiple goals can arise in practice.
12. Compare and contrast what-if analysis and goal seeking.
13. Describe the general process of simulation.
14. List some of the major advantages of simulation over optimization and vice versa.
15. Many computer games can be considered visual simulation. Explain why.
16. Explain why VIS is particularly helpful in implementing recommendations derived by computers.

Exercises

Teradata University Network (TUN) and Other Hands-on Exercises

1. Explore **teradatauniversitynetwork.com**, and determine how models are used in the BI cases and papers.
2. Create the spreadsheet models shown in Figures 8.3 and 8.4.
   a. What is the effect of a change in the interest rate from 8% to 10% in the spreadsheet model shown in Figure 8.3?
   b. For the original model in Figure 8.3, what interest rate is required to decrease the monthly payments by 20%? What change in the loan amount would have the same effect?
   c. In the spreadsheet shown in Figure 8.4, what is the effect of a prepayment of $200 per month? What prepayment would be necessary to pay off the loan in 25 years instead of 30 years?
3. Solve the MBI product-mix problem described in this chapter, using either Excel's Solver or a student version of an LP solver, such as Lindo. Lindo is available from Lindo Systems, Inc., at **lindo.com**; others are also available—search the Web. Examine the solution (output) reports for the answers and sensitivity report. Did you get the same results as reported in this chapter? Try the sensitivity analysis outlined in the chapter; that is, lower the right-hand side of the CC-8 marketing constraint by one unit, from 200 to 199. What happens to the solution when you solve this modified problem? Eliminate the CC-8 lower-bound constraint entirely (this can be done easily by either deleting it in Solver or setting the lower limit to zero) and re-solve the problem. What happens? Using the original formulation, try modifying the objective function coefficients and see what happens.
4. Investigate via a Web search how models and their solutions are used by the U.S. Department of Homeland Security in the "war against terrorism." Also investigate how other governments or government agencies are using models in their missions.
5. This problem was contributed by Dr. Rick Wilson of Oklahoma State University.

   The recent drought has hit farmers hard. Cows are eating candy corn!

   You are interested in creating a feed plan for the next week for your cattle using the following seven non-traditional feeding products: Chocolate Lucky Charms cereal, Butterfinger bars, Milk Duds, vanilla ice cream, Cap'n Crunch cereal, candy corn (because the real corn is all dead), and Chips Ahoy cookies.

	Choc Lucky Charms	Butterfinger	Milk Duds	Vanilla Ice Cream	Cap'n Crunch	Candy Corn	Chips Ahoy
$/lb	2.15	7	4.25	6.35	5.25	4	6.75
Choc	YES	YES	YES	NO	NO	NO	YES
Protein	75	80	45	65	72	26	62
TDN	12	20	18	6	11	8	12
Calcium	3	4	4.5	12	2	1	5

Their per pound cost is shown, as is the protein units per pound they contribute, the total digestible nutrients (TDN) they contribute per pound, and the calcium units per pound.

You estimate that the total amount of nontraditional feeding products contribute the following amount of nutrients: at least 20,000 units of protein, at least 4,025 units of TDN, at least 1,000 but no more than 1,200 units of calcium.

There are some other miscellaneous requirements as well.

- The chocolate in your overall feed plan (in pounds) cannot exceed the amount of nonchocolate poundage. Whether a product is considered chocolate

or not is shown in the table (YES = chocolate, NO = not chocolate).
- No one feeding product can make up more than 25% of the total pounds needed to create an acceptable feed mix.
- There are two cereals (Chocolate Lucky Charms and Cap'n Crunch). Combined, they can be no more than 40% (in pounds) of the total mix required to meet the mix requirements.

Determine the optimal levels of the seven products to create your weekly feed plan that minimizes cost. Note that all amounts of products must *not* have fractional values (whole numbered pounds only).

6. This exercise was also contributed by Dr. Rick Wilson of Oklahoma State University to illustrate the modeling capabilities of Excel Solver.

National signing day for rugby recruiting season 2018 has been completed. Now, as the recruiting coordinator for the San Diego State University Aztec rugby team, it is time to analyze the results and plan for 2019.

You've developed complex analytics and data collection processes and applied them for the past few recruiting seasons to help you develop a plan for 2019. Basically, you have divided the area in which you actively recruit rugby players into eight different regions. Each region has a per-target cost, a "star rating" (average recruit "star" ranking, from 0 to 5, similar to what Rivals uses for football), a yield or acceptance rate percentage (the percentage of targeted recruits who come to SDSU), and a visibility measure, which represents a measure of how much publicity SDSU gets for recruiting in that region, measured per target (increased visibility will enhance future recruiting efforts).

	Cost/target	avg star rating	acceptance rate %	visibility per target
Region1	125	3	40	0
Region2	89	2.5	42	0
Region3	234	3.25	25	2
Region4	148	3.1	30	3
Region5	321	3.5	22	7
Region6	274	3.45	20	4
Region7	412	3.76	17	5
Region8	326	3.2	18	5.5

Your goal is to create a LINEAR mathematical model that determines *the number of target recruits you should pursue in each region* in order to have an estimated yield (expected number) of *at least* 25 rugby recruits for next year while minimizing cost. (Region 1 with yield of 40%: if we target 10 people, the expected number that will come is .4*10 = 40.)

In determining the optimal number of targets in each region (which, not surprisingly, should be integer values), you must also satisfy the following conditions:
- No more than 20% of the total targets (not the expected number of recruits) should be from any one region.
- Each region should have at least 4% of the total targets (again, not the expected number of recruits, but the number of targets).
- The average star rating of the targets must be at least equal to 3.3.
- The average visibility value of the targets must be at least equal to 3.5.
- Off on the recruiting trail you go!

7. This exercise was also contributed by Dr. Rick Wilson of Oklahoma State University.

You are the Water Resources Manager for Thirstiville, OK, and are working out the details for next year's contracts with three different entities to supply water to your town. Each water source (A, B, C) provides water of different quality. The quality assessment is aggregated together in two values P1 and P2, representing a composite of contaminants, such as THMs, HAAs, and so on. The sources each have a maximum of water that they can provide (measured in thousands of gallons), a minimum that we must purchase from them, and a per-thousand-gallon cost.

	MIN	MAX	P1	P2	COST
Source A	400	1000	4	1	0.25
Source B	1000	2500	3.5	3	0.175
Source C	0	775	5	2.5	0.20

On the product end, you must procure water such that you can provide three distinct water products for next year (this is all being done at the aggregate "city" level). You must provide drinking water to the city, and then water to two different wholesale clients (this is commonly done by municipalities). The table below shows requirements for these three products, and the "sales" or revenue that you get from each customer (by thousand gallons, same scale as the earlier cost).

For each of the three water products/customers, MIN is the minimum that we have to provide to each, MAX is the maximum that we can provide (it is reasonable to be provided with a targeted range of product to provide to our customers), the maximum P1 and P2 weighted average for the water blended together for each quality "category" (the contaminants) per customer, and the sales price.

	MIN	MAX	P1	P2	SALES
Drinking	1500	1700	3.75	2.25	0.35
WSale 1	250	325	No Req.	2.75	0.4
WSale 2	No limit	No limit	4	2	0.425

Yes, the second wholesale customer (WSale 2) will take as much water as you can blend together for them.

Obviously, water from all three sources will need to be blended together to meet the Thirstiville customer requirements. There is one more requirement: for each of the three products (drinking water and the two wholesale clients), Source A and Source B both individually (yes, separately) must make up at least 20% of the

total amount of the production of that particular water type. We do not have such a requirement for Source C.

Create an appropriate LP model that determines how to meet customer water demand for next year *while maximizing profit (sales less costs)*. Summarize your results (something more than telepathy—say, some sort of table of data beyond the model solution?) It must use words (☺) and indicate how much water we should promise to buy from our three sources. Integers are not required.

References

"Canadian Football League Uses Frontline Solvers to Optimize Scheduling in 2016." **www.solver.com/news/canadian-football-league-uses-frontline-solvers-optimize-scheduling-2016** (accessed September 2018).

"Risk-Based Planning and Scheduling (RPS) with Simio." **Www.simio.com**, Simio LLC, **www.simio.com/about-simio/why-simio/simio-RPS-risk-based-planning-and-scheduling.php** (accessed September 2018).

Arsham, H. (2006a). "Modeling and Simulation Resources." **home.ubalt.edu/ntsbarsh/Business-stat/RefSim.htm** (accessed November 2018).

Arsham, H. (2006b). "Decision Science Resources." **home.ubalt.edu/ntsbarsh/Business-stat/Refop.htm** (accessed November 2018).

Arthur, Molly. "Simulation-Based Production Scheduling System." **www.simio.com**, Simio LLC, 2014, **www.simio.com/case-studies/A-Steel-Tubing-Manufacturer-Expects-More-Market-Share/A-Steel-Tubing-Manufacturer-Expects-More-Market-Share.pdf** (accessed September 2018).

Bailey, M. J., J. Snapp, S. Yetur, J. S. Stonebraker, S. A. Edwards, A. Davis, & R. Cox. (2011). "Practice Summaries: American Airlines Uses Should-Cost Modeling to Assess the Uncertainty of Bids for Its Full-Truckload Shipment Routes." *Interfaces, 41*(2), 194–196.

Banks, J., & Gibson, R. R. (2009). Seven Sins of Simulation Practice." *INFORMS Analytics*, 24–27. **www.analytics-magazine.org/summer-2009/193-strategic-problems-modeling-the-market-space** (accessed September 2018).

Bowers, M. R., C. E. Noon, W. Wu, & J. K. Bass. (2016). "Neonatal Physician Scheduling at the University of Tennessee Medical Center." *Interfaces, 46*(2), 168–182.

Chongwatpol, J., & R. Sharda. (2013). "RFID-Enabled Track and Traceability in Job-Shop Scheduling Environment." *European Journal of Operational Research, 227*(3), 453–463, **http://dx.doi.org/10.1016/j.ejor.2013.01.009**.

Christiansen, M., K. Fagerholt, G. Hasle, A. Minsaas, & B. Nygreen. (2009, April). "Maritime Transport Optimization: An Ocean of Opportunities." *OR/MS Today, 36*(2), 26–31.

Clemen, R. T., & Reilly, T. (2000). *Making Hard Decisions with Decision Tools Suite*. Belmont, MA: Duxbury Press.

Dilkina, B. N., & W. S. Havens. "The U.S. National Football League Scheduling Problem. Intelligent Systems Lab," **www.cs.cornell.edu/~bistra/papers/NFLsched1.pdf** (accessed September 2018).

Farasyn, I., K. Perkoz, & W. Van de Velde. (2008, July/August). "Spreadsheet Models for Inventory Target Setting at Procter & Gamble." *Interfaces, 38*(4), 241–250.

Goodwin, P., & Wright, G. (2000). *Decision Analysis for Management Judgment*, 2nd ed. New York: Wiley.

Hurley, W. J., & M. Balez. (2008, July/August). "A Spreadsheet Implementation of an Ammunition Requirements Planning Model for the Canadian Army." *Interfaces, 38*(4), 271–280.

ingrammicrocommerce.com, "CUSTOMERS," **https://www.ingrammicrocommerce.com/customers/** (accessed July 2016).

Kearns, G. S. (2004, January–March). "A Multi-Objective, Multicriteria Approach for Evaluating IT Investments: Results from Two Case Studies." *Information Resources Management Journal, 17*(1), 37–62.

Kelly, A. (2002). *Decision Making Using Game Theory: An Introduction for Managers*. Cambridge, UK: Cambridge University Press.

Knight, F. H. (1933). *Risk, Uncertainty and Profit: With an Additional Introductory Essay Hither to Unpublished*. London school of economics and political science.

Koksalan, M., & S. Zionts. (Eds.). (2001). *Multiple Criteria Decision Making in the New Millennium*. Berlin: Springer-Verlag.

Kontoghiorghes, E. J., B. Rustem, & S. Siokos. (2002). *Computational Methods in Decision Making, Economics, and Finance*. Boston: Kluwer.

Kostuk, Kent J., and K. A. Willoughby. (2012). "A Decision Support System for Scheduling the Canadian Football League." *Interfaces, 42*(3), 286–295.

Manikas, A. S., J. R. Kroes, & T. F. Gattiker. (2016). Metro Meals on Wheels Treasure Valley Employs a Low-Cost Routing Tool to Improve Deliveries. *Interfaces, 46*(2), 154–167.

Mookherjee, R., J. Martineau, L. Xu, M. Gullo, K. Zhou, A. Hazlewood, X. Zhang, F. Griarte, & N. Li. (2016). "End-to-End Predictive Analytics and Optimization in Ingram Micro's Two-Tier Distribution Business." *Interfaces, 46* (1),49–73.

Ovchinnikov, A., & J. Milner. (2008, July/August). "Spreadsheet Model Helps to Assign Medical Residents at the University of Vermont's College of Medicine." *Interfaces, 38*(4), 311–323.

Simio.com. "Cosan Case Study—Optimizing Agricultural LogisticsOperations." **http://www.simio.com/case-studies/Cosan-agricultural-logistics-simulation-software-case-study/index.php** (accessed September 2018).

Slaugh, V. W., M. Akan, O. Kesten, & M. U. Unver. (2016). "The Pennsylvania Adoption Exchange Improves Its Matching Process." *Interfaces*, 462, 133–154.

Solver.com. "Optimizing Vendor Contract Awards Gets an A+." **solver.com/news/optimizing-vendor-contract-awards-gets** (accessed September 2018).

Turban, E., & J. Meredith. (1994). *Fundamentals of Management Science*, 6th ed. Richard D. Irwin, Inc.

Wikipedia.com. Cosan. **https://en.wikipedia.org/wiki/Cosan** (accessed November 2018).

CHAPTER 9

Big Data, Cloud Computing, and Location Analytics: Concepts and Tools

LEARNING OBJECTIVES

- Learn what Big Data is and how it is changing the world of analytics
- Understand the motivation for and business drivers of Big Data analytics
- Become familiar with the wide range of enabling technologies for Big Data analytics
- Learn about Hadoop, MapReduce, and NoSQL as they relate to Big Data analytics
- Compare and contrast the complementary uses of data warehousing and Big Data technologies
- Become familiar with in-memory analytics and Spark applications
- Become familiar with select Big Data platforms and services
- Understand the need for and appreciate the capabilities of stream analytics
- Learn about the applications of stream analytics
- Describe the current and future use of cloud computing in business analytics
- Describe how geospatial and location-based analytics are assisting organizations

Big Data, which means many things to many people, is not a new technological fad. It has become a business priority that has the potential to profoundly change the competitive landscape in today's globally integrated economy. In addition to providing innovative solutions to enduring business challenges, Big Data and analytics instigate new ways to transform processes, organizations, entire industries, and even society altogether. Yet extensive media coverage makes it hard to distinguish hype from reality. This chapter aims to provide a comprehensive coverage of Big Data, its enabling technologies, and related analytics concepts to help understand the capabilities and limitations of this emerging technology. The chapter starts with a definition and related concepts of Big Data followed by the technical details of the enabling technologies, including Hadoop, MapReduce, and NoSQL. We provide a comparative analysis between data warehousing and Big Data analytics. The last part of the chapter is dedicated to stream

analytics, which is one of the most promising value propositions of Big Data analytics. This chapter contains the following sections:

9.1 OPENING VIGNETTE: Analyzing Customer Churn in a Telecom Company Using Big Data Methods

BACKGROUND

A telecom company (named Access Telecom [AT] for privacy reasons) wanted to stem the tide of customers churning from its telecom services. Customer churn in the telecommunications industry is common. However, Access Telecom was losing customers at an alarming rate. Several reasons and potential solutions were attributed to this phenomenon. The management of the company realized that many cancellations involved communications between the customer service department and the customers. To this end, a task force comprising members from the customer relations office and the information technology (IT) department was assembled to explore the problem further. Their task was to explore how the problem of customer churn could be reduced based on an analysis of the customers' communication patterns (Asamoah, Sharda, Zadeh, & Kalgotra, 2016).

BIG DATA HURDLES

Whenever a customer had a problem about issues such as their bill, plan, and call quality, they would contact the company in multiple ways. These included a call center, company Web site (contact us links), and physical service center walk-ins. Customers could cancel an account through one of these listed interactions. The company wanted to see if analyzing these customer interactions could yield any insights about the questions the customers asked or the contact channel(s) they used before canceling their account. The data generated because of these interactions were in both text and audio. So, AT would have to combine all the data into one location. The company explored the use of traditional platforms for data management but soon found they were not versatile enough to handle advanced data analysis in the scenario where there were multiple formats of data from multiple sources (Thusoo, Shao, & Anthony, 2010).

There were two major challenges in analyzing this data: multiple data sources leading to a variety of data and also a large volume of data.

1. **Data from multiple sources:** Customers could connect with the company by accessing their accounts on the company's Web site, allowing AT to generate Web log information on customer activity. The Web log track allowed the company to identify if and when a customer reviewed his/her current plan, submitted a complaint, or

checked the bill online. At the customer service center, customers could also lodge a service complaint, request a plan change, or cancel the service. These activities were logged into the company's transaction system and then the enterprise data warehouse. Last, a customer could call the customer service center on the phone and transact business just like he/she would do in person at a customer service center. Such transactions could involve a balance inquiry or an initiation of plan cancellation. Call logs were available in one system with a record of the reasons a customer was calling. For meaningful analysis to be performed, the individual data sets had to be converted into similar structured formats.

2. **Data volume:** The second challenge was the sheer quantity of data from the three sources that had to be extracted, cleaned, restructured, and analyzed. Although previous data analytics projects mostly utilized a small sample set of data for analysis, AT decided to leverage the multiple variety and sources of data as well as the large volume of data recorded to generate as many insights as possible.

An analytical approach that could make use of all the channels and sources of data, although huge, would have the potential of generating rich and in-depth insights from the data to help curb the churn.

SOLUTION

Teradata Vantage's unified Big Data architecture (previously offered as Teradata Aster) was utilized to manage and analyze the large multistructured data. We will introduce Teradata Vantage in Section 9.8. A schematic of which data was combined is shown in Figure 9.1. Based on each data source, three tables were created with each table containing the following variables: customer ID, channel of communication, date/time

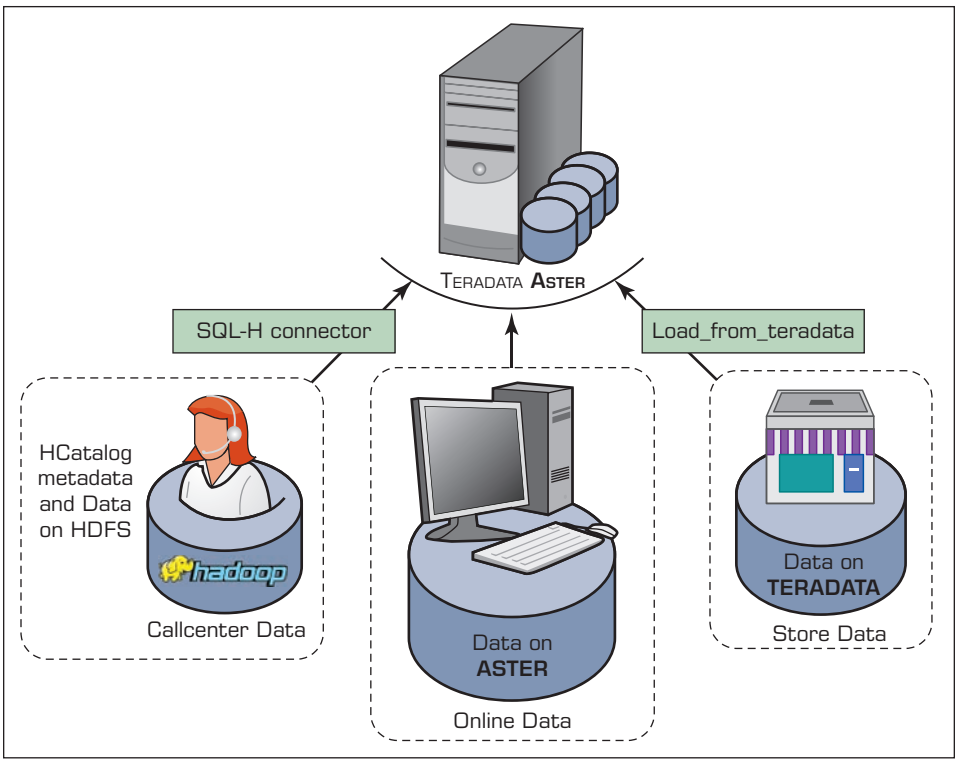


FIGURE 9.1 **Multiple Data Sources Integrated into Teradata Vantage.** *Source:* Teradata Corp.

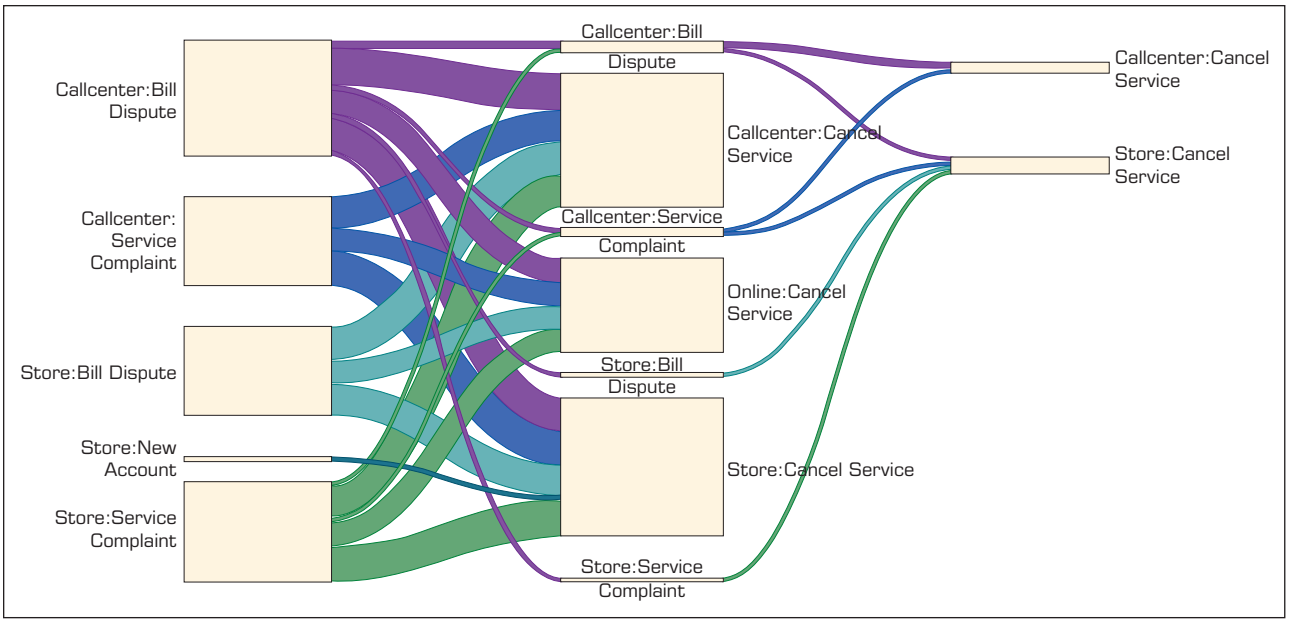


FIGURE 9.2 **Top 20 Paths Visualization.** *Source:* Teradata Corp.

stamp, and action taken. Prior to final cancellation of a service, the action-taken variable could be one or more of these 11 options (simplified for this case): present a bill dispute, request for plan upgrade, request for plan downgrade, perform profile update, view account summary, access customer support, view bill, review contract, access store locator function on the Web site, access frequently asked questions section on the Web site, or browse devices. The target of the analysis focused on finding the most common path resulting in a final service cancellation. The data was sessionized to group a string of events involving a particular customer into a defined time period (5 days over all the channels of communication) as one session. Finally, Vantage's nPath time sequence function (operationalized in an SQL-MapReduce framework) was used to analyze common trends that led to a cancellation.

RESULTS

The initial results identified several routes that could lead to a request for service cancellation. The company determined thousands of routes that a customer may take to cancel service. A follow-up analysis was performed to identify the most frequent routes to cancellation requests. This was termed as the Golden Path. The top 20 most occurring paths that led to a cancellation were identified in both short and long terms. A sample is shown in Figure 9.2.

This analysis helped the company identify a customer before they would cancel their service and offer incentives or at least escalate the problem resolution to a level where the customer's path to cancellation did not materialize.

QUESTIONS FOR THE OPENING VIGNETTE

1. What problem did customer service cancellation pose to AT's business survival?
2. Identify and explain the technical hurdles presented by the nature and characteristics of AT's data.
3. What is sessionizing? Why was it necessary for AT to sessionize its data?

4. Research other studies where customer churn models have been employed. What types of variables were used in those studies? How is this vignette different?
5. Besides Teradata Vantage, identify other popular Big Data analytics platforms that could handle the analysis described in the preceding case. (Hint: see Section 9.8.)

WHAT CAN WE LEARN FROM THIS VIGNETTE?

Not all business problems merit the use of a Big Data analytics platform. This situation presents a business case that warranted the use of a Big Data platform. The main challenge revolved around the characteristics of the data under consideration. The three different types of customer interaction data sets presented a challenge in analysis. The formats and fields of data generated in each of these systems was huge. And the volume was large as well. This made it imperative to use a platform that uses technologies to permit analysis of a large volume of data that comes in a variety of formats.

Recently, Teradata stopped marketing Aster as a separate product and has merged all of the Aster capabilities into its new offering called Teradata Vantage. Although that change somewhat impacts how the application would be developed today, it is still a terrific example of how a variety of data can be brought together to make business decisions.

It is also worthwhile to note that AT aligned the questions asked of the data with the organization's business strategy. The questions also informed the type of analysis that was performed. It is important to understand that for any application of a Big Data architecture, the organization's business strategy and the generation of relevant questions are key to identifying the type of analysis to perform.

Sources: D. Asamoah, R. Sharda, A. Zadeh, & P. Kalgotra. (2016). "Preparing a Big Data Analytics Professional: A Pedagogic Experience." In *DSI 2016 Conference*, Austin, TX. A. Thusoo, Z. Shao, & S. Anthony. (2010). "Data Warehousing and Analytics Infrastructure at Facebook." In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* (p. 1013). doi: 10.1145/1807167.1807278.

9.2 DEFINITION OF BIG DATA

Using data to understand customers/clients and business operations to sustain (and foster) growth and profitability is an increasingly challenging task for today's enterprises. As more and more data becomes available in various forms and fashions, timely processing of the data with traditional means becomes impractical. Nowadays, this phenomenon is called Big Data, which is receiving substantial press coverage and drawing increasing interest from both business users and IT professionals. The result is that Big Data is becoming an overhyped and overused marketing buzzword, leading some industry experts to argue dropping this phrase altogether.

Big Data means different things to people with different backgrounds and interests. Traditionally, the term *Big Data* has been used to describe the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA. But for most businesses, it's a relative term: "Big" depends on an organization's size. The point is more about finding new value within and outside conventional data sources. Pushing the boundaries of data analytics uncovers new insights and opportunities, and "big" depends on where you start and how you proceed. Consider the popular description of Big Data: Big Data exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span for its user population. **Big Data** has become a popular term to describe the exponential growth, availability, and use of information, both structured and unstructured. Much has been written on the Big Data trend and how it can serve as the basis for innovation,

differentiation, and growth. Because of the technology challenges in managing the large volume of data coming from multiple sources, sometimes at a rapid speed, additional new technologies have been developed to overcome the technology challenges. Use of the term *Big Data* is usually associated with such technologies. Because a prime use of storing such data is generating insights through analytics, sometimes the term Big Data is expanded as Big Data analytics. But the term is becoming content free in that it can mean different things to different people. Because our goal is to introduce you to the large data sets and their potential in generating insights, we will use the original term in this chapter.

Where does Big Data come from? A simple answer is "everywhere." The sources that were ignored because of the technical limitations are now treated as gold mines. Big Data may come from Web logs, radio-frequency identification (RFID), global positioning systems (GPS), sensor networks, social networks, Internet-based text documents, Internet search indexes, detail call records, astronomy, atmospheric science, biology, genomics, nuclear physics, biochemical experiments, medical records, scientific research, military surveillance, photography archives, video archives, and large-scale e-commerce practices.

Big Data is not new. What is new is that the definition and the structure of Big Data constantly change. Companies have been storing and analyzing large volumes of data since the advent of the data warehouses in the early 1990s. Whereas terabytes used to be synonymous with Big Data warehouses, now it's exabytes, and the rate of growth in data volume continues to escalate as organizations seek to store and analyze greater levels of transaction details, as well as Web- and machine-generated data, to gain a better understanding of customer behavior and business drivers.

Many (academics and industry analysts/leaders alike) think that "Big Data" is a misnomer. What it says and what it means are not exactly the same. That is, Big Data is not just "big." The sheer volume of the data is only one of many characteristics that are often associated with Big Data, including variety, velocity, veracity, variability, and value proposition, among others.

The "V"s That Define Big Data

Big Data is typically defined by three "V"s: volume, variety, velocity. In addition to these three, we see some of the leading Big Data solution providers adding other "V"s, such as veracity (IBM), variability (SAS), and value proposition.

VOLUME Volume is obviously the most common trait of Big Data. Many factors contributed to the exponential increase in data volume, such as transaction-based data stored through the years, text data constantly streaming in from social media, increasing amounts of sensor data being collected, automatically generated RFID and GPS data, and so on. In the past, excessive data volume created storage issues, both technical and financial. But with today's advanced technologies coupled with decreasing storage costs, these issues are no longer significant; instead, other issues have emerged, including how to determine relevance amid the large volumes of data and how to create value from data that is deemed to be relevant.

As mentioned before, *big* is a relative term. It changes over time and is perceived differently by different organizations. With the staggering increase in data volume, even the naming of the next Big Data echelon has been a challenge. The highest mass of data that used to be called petabytes (PB) has left its place to zettabytes (ZB), which is a trillion gigabytes (GB) or a billion terabytes (TB). Technology Insights 9.1 provides an overview of the size and naming of Big Data volumes.

From a short historical perspective, in 2009 the world had about 0.8 ZB of data; in 2010, it exceeded the 1 ZB mark; at the end of 2011, the number was 1.8 ZB. It is expected to be 44 ZB in 2020 (Adshead, 2014). With the growth of sensors and the Internet of Things (IoT—to be introduced in the next chapter), these forecasts could all be wrong.

Though these numbers are astonishing in size, so are the challenges and opportunities that come with them.

VARIETY Data today come in all types of formats—ranging from traditional databases to hierarchical data stores created by the end users and OLAP systems to text documents, e-mail, XML, meter-collected and sensor-captured data, to video, audio, and stock ticker data. By some estimates, 80 to 85% of all organizations' data are in some sort of unstructured or semi-structured format (a format that is not suitable for traditional database schemas). But there is no denying its value, and hence, it must be included in the analyses to support decision making.

VELOCITY According to Gartner, velocity means both how fast data is being produced and how fast the data must be processed (i.e., captured, stored, and analyzed) to meet the need or demand. RFID tags, automated sensors, GPS devices, and smart meters are driving an increasing need to deal with torrents of data in near real time. Velocity is perhaps the most overlooked characteristic of Big Data. Reacting quickly enough to deal with velocity is a challenge to most organizations. For time-sensitive environments, the opportunity cost clock of the data starts ticking the moment the data is created. As time passes, the value proposition of the data degrades and eventually becomes worthless. Whether the subject matter is the health of a patient, the well-being of a traffic system, or the health of an investment portfolio, accessing the data and reacting faster to the circumstances will always create more advantageous outcomes.

TECHNOLOGY INSIGHTS 9.1 The Data Size Is Getting Big, Bigger, and Bigger

The measure of data size is having a hard time keeping up with new names. We all know kilobyte (KB, which is 1,000 bytes), megabyte (MB, which is 1,000,000 bytes), gigabyte (GB, which is 1,000,000,000 bytes), and terabyte (TB, which is 1,000,000,000,000 bytes). Beyond that, the names given to data sizes are relatively new to most of us. The following table shows what comes after terabyte and beyond.

Name	Symbol	Value
Kilobyte	kB	10^3
Megabyte	MB	10^6
Gigabyte	GB	10^9
Terabyte	TB	10^{12}
Petabyte	PB	10^{15}
Exabyte	EB	10^{18}
Zettabyte	ZB	10^{21}
Yottabyte	YB	10^{24}
Brontobyte*	BB	10^{27}
Gegobyte*	GeB	10^{30}

*Not an official SI (International System of Units) name/symbol, yet.

Consider that an exabyte of data is created on the Internet each day, which equates to 250 million DVDs' worth of information. And the idea of even larger amounts of data—a zettabyte—isn't too far off when it comes to the amount of information traversing the Web in any one year. In fact, industry experts are already estimating that we will see 1.3 zettabytes of traffic annually

over the Internet by 2016—and it could jump to 2.3 zettabytes by 2020. By 2020, Internet traffic is expected to reach 300 GB per capita per year. When referring to yottabytes, some of the Big Data scientists often wonder about how much data the NSA or FBI have on people altogether. Put in terms of DVDs, a yottabyte would require 250 trillion of them. A brontobyte, which is not an official SI prefix but is apparently recognized by some people in the measurement community, is a 1 followed by 27 zeros. The size of such a magnitude can be used to describe the amount of sensor data that we will get from the Internet in the next decade, if not sooner.

A gegobyte is 10 to the power of 30. With respect to where the Big Data comes from, consider the following:

- The CERN Large Hadron Collider generates 1 petabyte per second.
- Sensors from a Boeing jet engine create 20 terabytes of data every hour.
- Every day, 600 terabytes of new data are ingested in Facebook databases.
- On YouTube, 300 hours of video are uploaded per minute, translating to 1 terabyte every minute.
- The proposed Square Kilometer Array telescope (the world's proposed biggest telescope) will generate an exabyte of data per day.

Sources: S. Higginbotham. (2012). "As Data Gets Bigger, What Comes after a Yottabyte?" **gigaom.com/2012/10/30/as-data-gets-bigger-what-comes-after-a-yottabyte** (accessed October 2018). Cisco. (2016). "The Zettabyte Era: Trends and Analysis." **cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf** (accessed October 2018).

In the Big Data storm that we are currently witnessing, almost everyone is fixated on at-rest analytics, using optimized software and hardware systems to mine large quantities of variant data sources. Although this is critically important and highly valuable, there is another class of analytics, driven from the velocity of Big Data, called "data stream analytics" or "in-motion analytics," which is evolving fast. If done correctly, data stream analytics can be as valuable as, and in some business environments more valuable than at-rest analytics. Later in this chapter we will cover this topic in more detail.

VERACITY *Veracity* is a term coined by IBM that is being used as the fourth "V" to describe Big Data. It refers to conformity to facts: accuracy, quality, truthfulness, or trustworthiness of the data. Tools and techniques are often used to handle Big Data's veracity by transforming the data into quality and trustworthy insights.

VARIABILITY In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something big trending in the social media? Perhaps there is a high-profile IPO looming. Maybe swimming with pigs in the Bahamas is suddenly the must-do vacation activity. Daily, seasonal, and event-triggered peak data loads can be highly variable and thus challenging to manage—especially with social media involved.

VALUE PROPOSITION The excitement around Big Data is its value proposition. A preconceived notion about "Big" data is that it contains (or has a greater potential to contain) more patterns and interesting anomalies than "small" data. Thus, by analyzing large and feature-rich data, organizations can gain greater business value that they may not have otherwise. Although users can detect the patterns in small data sets using simple statistical and machine-learning methods or ad hoc query and reporting tools, Big Data means "big" analytics. Big analytics means greater insight and better decisions, something that every organization needs.

Because the exact definition of Big Data (or its successor terms) is still a matter of ongoing discussion in academic and industrial circles, it is likely that more characteristics (perhaps more "V"s) are likely to be added to this list. Regardless of what happens, the importance and value proposition of Big Data is here to stay. Figure 9.3 shows a conceptual architecture where Big Data (at the left side of the figure) is converted to business

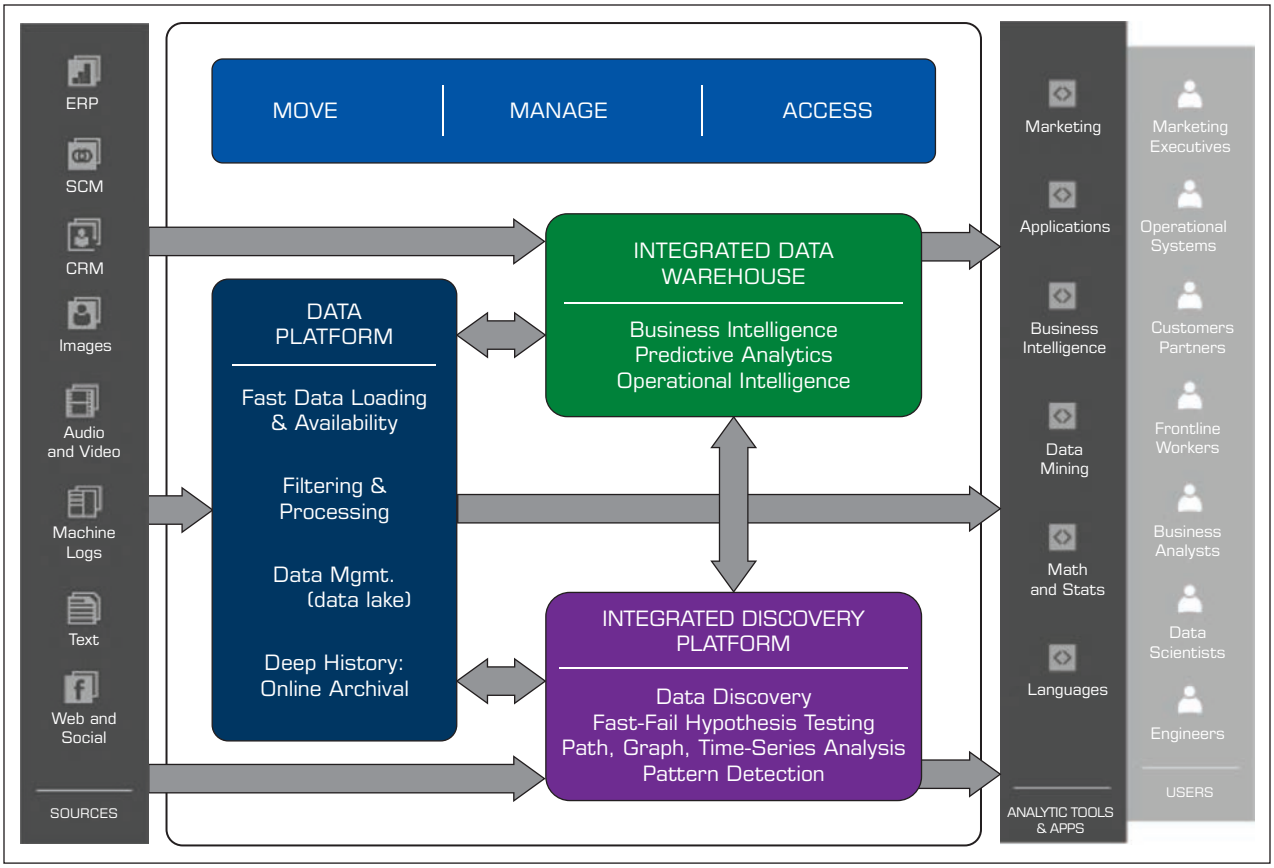


FIGURE 9.3 **A High-Level Conceptual Architecture for Big Data Solutions.** *Source:* Teradata Company.

insight through the use of a combination of advanced analytics and delivered to a variety of different users/roles for faster/better decision making.

Another term that is being added to Big Data buzzwords is alternative data. Application Case 9.1 shows examples of multiple types of data in a number of different scenarios.

Application Case 9.1 Alternative Data for Market Analysis or Forecasts

Getting a good forecast and understanding of the situation is crucial for any scenario, but it is especially important to players in the investment industry. Being able to get an early indication of how a particular retailer's sales are doing can give an investor a leg up on whether to buy or sell that retailer's stock even before the earnings reports are released. The problem of forecasting economic activity or microclimates based on a variety of data beyond the usual retail data is a very recent phenomenon and has led to another buzzword—"alternative data." A major mix in this alternative data category is satellite imagery, but it also includes other data such as social media, government filings, job postings, traffic patterns, changes in parking lots or open spaces detected by satellite images, mobile phone usage patterns in any given location at any given time, search patterns on search engines, and so on. Facebook and other companies have invested in satellites to try to image the whole globe every day so that daily changes can be tracked at any location

(*Continued*)

Application Case 9.1 (Continued)

and the information can be used for forecasting. Many interesting examples of more reliable and advanced forecasts have been reported. Indeed, this activity is being led by start-up companies. Tartar (2018) cited several examples. We mentioned some in Chapter 1. Here are some of the examples identified by them and many other proponents of alternative data:

- RS Metrics monitored parking lots across the United States for various hedge funds. In 2015, based on an analysis of the parking lots, RS Metrics predicted a strong second quarter in 2015 for JC Penney. Its clients (mostly hedge funds) profited from this advanced insight. A similar story has been reported for Wal-Mart using car counts in its parking lots to forecast sales.
- Spaceknow keeps track of changes in factory surroundings for over 6,000 Chinese factory sites. Using this data, the company has been able to provide a better idea of China's industrial economic activity than what the Chinese government has been reporting.
- Telluslabs, Inc. compiles data from NASA and European satellites to build prediction models for various crops such as corn, rice, soybean, wheat, and so on. Besides the images from the satellites, they incorporate measurements of thermal infrared bands, which help measure radiating heat to predict health of the crops.
- DigitalGlobe is able to analyze the size of a forest with more accuracy because its software can count every single tree in a forest. This results in a more accurate estimate because there is no need to use a representative sample.

These examples illustrate just a sample of ways data can be combined to generate new insights. Of course, there are privacy concerns in some cases. For example, Yodlee, a division of Envestnet, provides personal finance tools to many large banks as well as personal financial tools to individuals. Thus, it has access to massive information about individuals. It has faced concerns about the privacy and security of this information, especially in light of the major breaches reported by Facebook, Cambridge Analytics, and Equifax. Although such concerns will eventually be resolved by policy makers or the market, what is clear is that new and interesting ways of combining satellite data and many other data sources are spawning a new crop of analytics companies. All of these organizations are working with data that meets the three V's—variety, volume, and velocity characterizations. Some of these companies also work with another category of data—sensors. But this group of companies certainly also falls under a group of innovative and emerging applications.

Sources: C. Dillow. (2016). "What Happens When You Combine Artificial Intelligence and Satellite Imagery." **fortune.com/2016/03/30/facebook-ai-satellite-imagery/** (accessed October 2018). G. Ekster. (2015). "Driving Investment Performance with Alternative Data." **integrity-research.com/wp-content/uploads/2015/11/Driving-Investment-Performance-With-Alternative-Data.pdf** (accessed October 2018). B. Hope. (2015). "Provider of Personal Finance Tools Tracks Bank Cards, Sells Data to Investors." **wsj.com/articles/provider-of-personal-finance-tools-tracks-bank-cards-sells-data-to-investors-1438914620** (accessed October 2018). C. Shaw. (2016). "Satellite Companies Moving Markets." **quandl.com/blog/alternative-data-satellite-companies** (accessed October 2018). C. Steiner. (2009). "Sky High Tips for Crop Traders." **www.forbes.com/forbes/2009/0907/technology-software-satellites-sky-high-tips-for-crop-traders.html** (accessed October 2018). M. Turner. (2015). "This Is the Future of Investing, and You Probably Can't Afford It." **businessinsider.com/hedge-funds-are-analysing-data-to-get-an-edge-2015-8** (accessed October 2018).

Questions for Discussion

1. What is a common thread in the examples discussed in this application case?
2. Can you think of other data streams that might help give an early indication of sales at a retailer?
3. Can you think of other applications along the lines presented in this application case?

SECTION 9.2 REVIEW QUESTIONS

1. Why is Big Data important? What has changed to put it in the center of the analytics world?
2. How do you define Big Data? Why is it difficult to define?
3. Out of the "V"s that are used to define Big Data, in your opinion, which one is the most important? Why?
4. What do you think the future of Big Data will be like? Will it leave its popularity to something else? If so, what will it be?

9.3 FUNDAMENTALS OF BIG DATA ANALYTICS

Big Data by itself, regardless of the size, type, or speed, is worthless unless business users do something with it that delivers value to their organizations. That's where "big" analytics comes into the picture. Although organizations have always run reports and dashboards against data warehouses, most have not opened these repositories to in-depth on-demand exploration. This is partly because analysis tools are too complex for the average user but also because the repositories often do not contain all the data needed by the power user. But this is about to change (and has been changing, for some) in a dramatic fashion, thanks to the new Big Data analytics paradigm.

With the value proposition, Big Data also brought about big challenges for organizations. The traditional means for capturing, storing, and analyzing data are not capable of dealing with Big Data effectively and efficiently. Therefore, new breeds of technologies need to be developed (or purchased/hired/outsourced) to take on the Big Data challenge. Before making such an investment, organizations should justify the means. Here are some questions that may help shed light on this situation. If any of the following statements are true, then you need to seriously consider embarking on a Big Data journey.

- You can't process the amount of data that you want to because of the limitations posed by your current platform or environment.
- You want to involve new/contemporary data sources (e.g., social media, RFID, sensory, Web, GPS, textual data) into your analytics platform, but you can't because it does not comply with the data storage schema-defined rows and columns without sacrificing fidelity or the richness of the new data.
- You need to (or want to) integrate data as quickly as possible to be current on your analysis.
- You want to work with a schema-on-demand (as opposed to predetermined schema used in relational database management systems [RDBMSs]) data storage paradigm because the nature of the new data may not be known, or there may not be enough time to determine it and develop a schema for it.
- The data is arriving so fast at your organization's doorstep that your traditional analytics platform cannot handle it.

As is the case with any other large IT investment, the success in Big Data analytics depends on a number of factors. Figure 9.4 shows a graphical depiction of the most critical success factors (Watson, 2012).

The following are the most critical success factors for **Big Data analytics** (Watson, Sharda, & Schrader, 2012):

1. ***A clear business need (alignment with the vision and the strategy).*** Business investments ought to be made for the good of the business, not for the sake of mere technology advancements. Therefore, the main driver for Big Data analytics should be the needs of the business, at any level—strategic, tactical, and operations.
2. ***Strong, committed sponsorship (executive champion).*** It is a well-known fact that if you don't have strong, committed executive sponsorship, it is difficult (if not impossible) to succeed. If the scope is a single or a few analytical applications, the sponsorship can be at the departmental level. However, if the target is enterprise-wide organizational transformation, which is often the case for Big Data initiatives, sponsorship needs to be at the highest levels and organization wide.
3. ***Alignment between the business and IT strategy.*** It is essential to make sure that the analytics work is always supporting the business strategy, and not the other way around. Analytics should play the enabling role in successfully executing the business strategy.

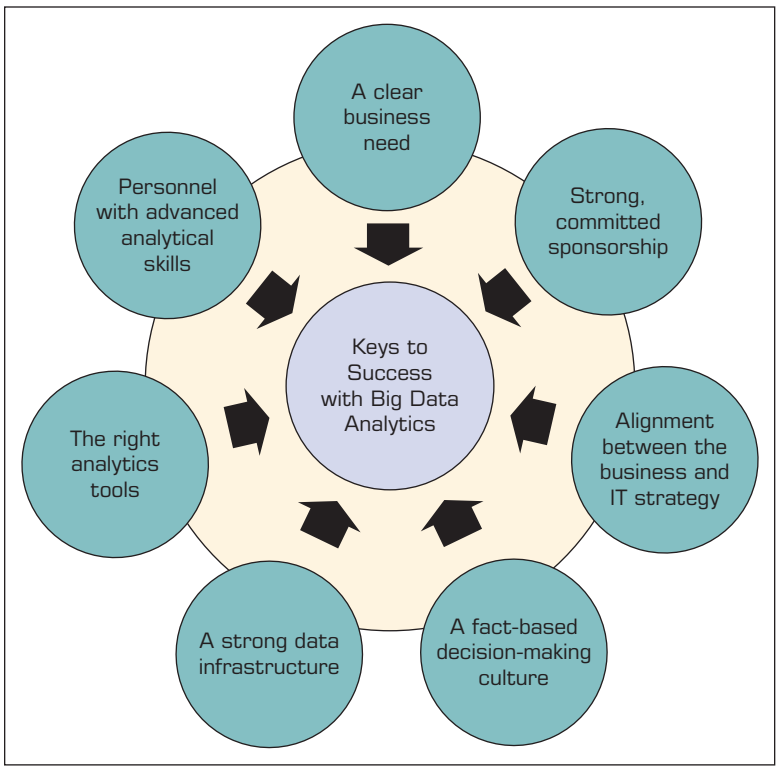


FIGURE 9.4 Critical Success Factors for Big Data Analytics. *Source:* Watson, H. (2012). The requirements for being an analytics-based organization. Business Intelligence Journal, 17(2), 42–44.

4. ***A fact-based decision-making culture.*** In a fact-based decision-making culture, the numbers rather than intuition, gut feeling, or supposition drive decision making. There is also a culture of experimentation to see what works and what doesn't. To create a fact-based decision-making culture, senior management needs to:
   - Recognize that some people can't or won't adjust
   - Be a vocal supporter
   - Stress that outdated methods must be discontinued
   - Ask to see what analytics went into decisions
   - Link incentives and compensation to desired behaviors
5. ***A strong data infrastructure.*** Data warehouses have provided the data infrastructure for analytics. This infrastructure is changing and being enhanced in the Big Data era with new technologies. Success requires marrying the old with the new for a holistic infrastructure that works synergistically.

As the size and complexity increase, the need for more efficient analytical systems is also increasing. To keep up with the computational needs of Big Data, a number of new and innovative computational techniques and platforms have been developed. These techniques are collectively called *high-performance computing*, which includes the following:

- ***In-memory analytics:*** Solves complex problems in near real time with highly accurate insights by allowing analytical computations and Big Data to be processed in-memory and distributed across a dedicated set of nodes.
- ***In-database analytics:*** Speeds time to insights and enables better data governance by performing data integration and analytic functions inside the database so you won't have to move or convert data repeatedly.

- ***Grid computing:*** Promotes efficiency, lower cost, and better performance by processing jobs in a shared, centrally managed pool of IT resources.
- ***Appliances:*** Brings together hardware and software in a physical unit that is not only fast but also scalable on an as-needed basis.

Computational requirements are just a small part of the list of challenges that Big Data impose on today's enterprises. The following is a list of challenges that are found by business executives to have a significant impact on successful implementation of Big Data analytics. When considering Big Data projects and architecture, being mindful of these challenges will make the journey to analytics competency a less stressful one.

Data volume: The ability to capture, store, and process a huge volume of data at an acceptable speed so that the latest information is available to decision makers when they need it.

Data integration: The ability to combine data that is not similar in structure or source and to do so quickly and at a reasonable cost.

Processing capabilities: The ability to process data quickly, as it is captured. The traditional way of collecting and processing data may not work. In many situations, data needs to be analyzed as soon as it is captured to leverage the most value. (This is called *stream analytics*, which will be covered later in this chapter.)

Data governance: The ability to keep up with the security, privacy, ownership, and quality issues of Big Data. As the volume, variety (format and source), and velocity of data change, so should the capabilities of governance practices.

Skills availability: Big Data is being harnessed with new tools and is being looked at in different ways. There is a shortage of people (often called *data scientists*) with skills to do the job.

Solution cost: Because Big Data has opened up a world of possible business improvements, a great deal of experimentation and discovery is taking place to determine the patterns that matter and the insights that turn to value. To ensure a positive return on investment on a Big Data project, therefore, it is crucial to reduce the cost of the solutions used to find that value.

Though the challenges are real, so is the value proposition of Big Data analytics. Anything that you can do as a business analytics leader to help prove the value of new data sources to the business will move your organization beyond experimenting and exploring Big Data into adapting and embracing it as a differentiator. There is nothing wrong with exploration, but ultimately the value comes from putting those insights into action.

Business Problems Addressed by Big Data Analytics

The top business problems addressed by Big Data overall are process efficiency and cost reduction, as well as enhancing customer experience, but different priorities emerge when it is looked at by industry. Process efficiency and cost reduction are perhaps among the top-ranked problems that can be addressed with Big Data analytics for the manufacturing, government, energy and utilities, communications and media, transport, and healthcare sectors. Enhanced customer experience may be at the top of the list of problems addressed by insurance companies and retailers. Risk management usually is at the top of the list for companies in banking and education. Here is a partial list of problems that can be addressed using Big Data analytics:

Process efficiency and cost reduction
Brand management
Revenue maximization, cross-selling, and up-selling
Enhanced customer experience
Churn identification, customer recruiting

Improved customer service
Identifying new products and market opportunities
Risk management
Regulatory compliance
Enhanced security capabilities

Application Case 9.2 illustrates an excellent example in the retail industry, where disparate data sources are integrated into a Big Data infrastructure to understand customer journeys.

Application Case 9.2 Overstock.com Combines Multiple Datasets to Understand Customer Journeys

Major retail organizations such as **Overstock.com** invest in many marketing campaigns to grow their revenue. These may include targeted online and direct mail campaigns, advertising through various channels, growing the loyalty program by providing different customer incentives, and so on. Each of these entails significant marketing costs and yet has varying levels of ROI. A challenge for any company analyzing all these campaigns is to unify these data in one location and understand customer journeys. Which combinations of campaigns or interactions eventually led to customers purchasing some items, and at what level? These data sources may include Web site traffic data that are in somewhat unstructured log files; e-mail campaign performance data may come through e-mail campaign companies in semistructured format; social media data such as Facebook posts and responses are in yet different streams. Linking all of these data to company's internal product data to assign value to a customer's purchase to be able to compute ROI on the combinations of campaigns is another data integration challenge. But combining such data sources is more practical under the Big Data framework. Then by using Path analysis capabilities that were illustrated in the opening vignette (Section 9.1), a nontechnical user can also look at various customer journeys and identify the ones that lead to most efficient sales and a high ROI for marketing efforts. The goal is to build a long-term relationship with the customers by understanding their search patterns, purchase behaviors, website responses, and so on. **Overstock.com** has been able to achieve this successfully using Teradata Vantage's path analysis functionality but by combining the very different data sources under the Big data framework.

Questions for Discussion:

1. What are some of the different marketing campaigns a company might run to woo customers? What format might data about these campaigns take?
2. By visualizing the most common customer paths to sales, how would you use that information to make decisions on the future marketing campaigns?
3. What other applications of such path analysis techniques can you think of?

Compiled from: "Overstock.com Uses Teradata Path Analysis to Boost Its Customer Journey Analytics," March 27, 2018, at **www.retailitinsights.com/doc/overstock-com-uses-teradata-path-analysis-boost-customer-journey-analytics-0001** (accessed October 2018), and "Overstock.com: Revolutionizing Data and Analytics to Connect Soulfully with Their Customers," at **www.teradata.com/Resources/Videos/Overstock-com-Revolutionizing-data-and-analy** (accessed October 2018).

This section has introduced the basics of Big Data and some potential applications. In the next section we will learn about a few terms and technologies that have emerged in Big Data space.

SECTION 9.3 REVIEW QUESTIONS

1. What is Big Data analytics? How does it differ from regular analytics?
2. What are the critical success factors for Big Data analytics?
3. What are the big challenges that one should be mindful of when considering implementation of Big Data analytics?
4. What are the common business problems addressed by Big Data analytics?

9.4 BIG DATA TECHNOLOGIES

There are a number of technologies for processing and analyzing Big Data, but most have some common characteristics (Kelly, 2012). Namely, they take advantage of commodity hardware to enable scale-out and parallel-processing techniques; employ nonrelational data storage capabilities to process unstructured and semistructured data; and apply advanced analytics and data visualization technology to Big Data to convey insights to end users. The three Big Data technologies that stand out that most believe will transform the business analytics and data management markets are MapReduce, Hadoop, and NoSQL.

MapReduce

MapReduce is a technique popularized by Google that distributes the processing of very large multistructured data files across a large cluster of machines. High performance is achieved by breaking the processing into small units of work that can be run in parallel across the hundreds, potentially thousands, of nodes in the cluster. To quote the seminal paper on MapReduce:

> MapReduce is a programming model and an associated implementation for processing and generating large data sets. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. (Dean & Ghemawat, 2004)

The key point to note from this quote is that MapReduce is a programming model, not a programming language, that is, it is designed to be used by programmers, rather than business users. The easiest way to describe how MapReduce works is through the use of an example (see the Colored Square Counter in Figure 9.5).

The input to the MapReduce process in Figure 9.5 is a set of colored squares. The objective is to count the number of squares of each color. The programmer in this example is responsible for coding the map and reducing programs; the remainder of the processing is handled by the software system implementing the MapReduce programming model.

The MapReduce system first reads the input file and splits it into multiple pieces. In this example, there are two splits, but in a real-life scenario, the number of splits would typically be much higher. These splits are then processed by multiple map programs running in parallel on the nodes of the cluster. The role of each map program in this case is to group the data in a split by color. The MapReduce system then takes the output from each map program and merges (shuffle/sort) the results for input to the reduce program, which calculates the sum of the number of squares of each color. In this example, only one copy of the reduce program is used, but there may be more in practice. To optimize performance, programmers can provide their own shuffle/sort program and can also deploy a combiner that combines local map output files to reduce the number of output files that have to be remotely accessed across the cluster by the shuffle/sort step.

Why Use MapReduce?

MapReduce aids organizations in processing and analyzing large volumes of multistructured data. Application examples include indexing and search, graph analysis, text analysis, machine learning, data transformation, and so forth. These types of applications are often difficult to implement using the standard SQL employed by relational DBMSs.

The procedural nature of MapReduce makes it easily understood by skilled programmers. It also has the advantage that developers do not have to be concerned with implementing parallel computing—this is handled transparently by the system. Although

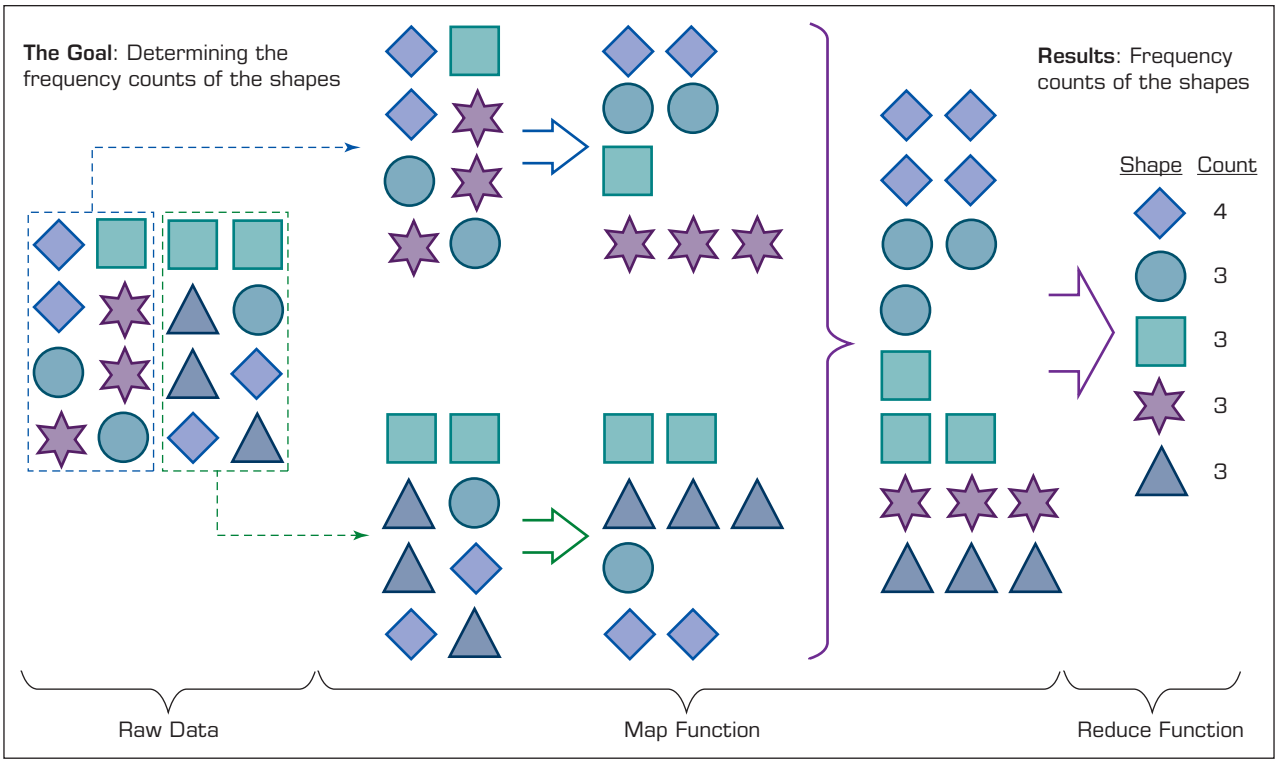


FIGURE 9.5 **A Graphical Depiction of the MapReduce Process.**

MapReduce is designed for programmers, nonprogrammers can exploit the value of pre-built MapReduce applications and function libraries. Both commercial and open source MapReduce libraries are available that provide a wide range of analytic capabilities. Apache Mahout, for example, is an open source machine-learning library of "algorithms for clustering, classification and batch-based collaborative filtering" that are implemented using MapReduce.

Hadoop

Hadoop is an open source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data. Originally created by Doug Cutting at Yahoo!, Hadoop was inspired by MapReduce, a user-defined function developed by Google in the early 2000s for indexing the Web. It was designed to handle petabytes and exabytes of data distributed over multiple nodes in parallel. Hadoop clusters run on inexpensive commodity hardware so projects can scale-out without breaking the bank. Hadoop is

Courtesy of The Apache Software Foundation.

now a project of the Apache Software Foundation, where hundreds of contributors continuously improve the core technology. Fundamental concept: Rather than banging away at one huge block of data with a single machine, Hadoop breaks up Big Data into multiple parts so each part can be processed and analyzed at the same time.

How Does Hadoop Work?

A client accesses unstructured and semistructured data from sources including log files, social media feeds, and internal data stores. It breaks the data up into "parts," which are then loaded into a file system made up of multiple nodes running on commodity hardware. The default file store in Hadoop is the **Hadoop Distributed File System, or HDFS**. File systems such as HDFS are adept at storing large volumes of unstructured and semistructured data as they do not require data to be organized into relational rows and columns. Each "part" is replicated multiple times and loaded into the file system so that if a node fails, another node has a copy of the data contained on the failed node. A Name Node acts as facilitator, communicating back to the client information such as which nodes are available, where in the cluster certain data resides, and which nodes have failed.

Once the data is loaded into the cluster, it is ready to be analyzed via the MapReduce framework. The client submits a "Map" job—usually a query written in Java—to one of the nodes in the cluster known as the Job Tracker. The Job Tracker refers to the Name Node to determine which data it needs to access to complete the job and where in the cluster that data is located. Once determined, the Job Tracker submits the query to the relevant nodes. Rather than bringing all the data back into a central location for processing, the processing occurs at each node simultaneously, or in parallel. This is an essential characteristic of Hadoop.

When each node has finished processing its given job, it stores the results. The client initiates a "Reduce" job through the Job Tracker in which results of the map phase stored locally on individual nodes are aggregated to determine the "answer" to the original query, and then are loaded onto another node in the cluster. The client accesses these results, which can then be loaded into one of a number of analytic environments for analysis. The MapReduce job has now been completed.

Once the MapReduce phase is complete, the processed data is ready for further analysis by data scientists and others with advanced data analytics skills. **Data scientists** can manipulate and analyze the data using any of a number of tools for any number of uses, including searching for hidden insights and patterns, or use as the foundation for building user-facing analytic applications. The data can also be modeled and transferred from Hadoop clusters into existing relational databases, data warehouses, and other traditional IT systems for further analysis and/or to support transactional processing.

Hadoop Technical Components

A Hadoop "stack" is made up of a number of components, which include

- ***Hadoop Distributed File System (HDFS):*** The default storage layer in any given Hadoop cluster.
- ***Name Node:*** The node in a Hadoop cluster that provides the client information on where in the cluster particular data is stored and if any nodes fail.
- ***Secondary Node:*** A backup to the Name Node, it periodically replicates and stores data from the Name Node should it fail.
- ***Job Tracker:*** The node in a Hadoop cluster that initiates and coordinates MapReduce jobs or the processing of the data.
- ***Slave Nodes:*** The grunts of any Hadoop cluster, slave nodes store data and take direction to process it from the Job Tracker.

In addition to these components, the Hadoop ecosystem is made up of a number of complementary subprojects. NoSQL data stores like Cassandra and HBase are also used to store the results of MapReduce jobs in Hadoop. In addition to Java, some MapReduce jobs and other Hadoop functions are written in **Pig**, an open source language designed specifically for Hadoop. **Hive** is an open source data warehouse originally developed by Facebook that allows for analytic modeling within Hadoop. Here are the most commonly referenced subprojects for Hadoop.

HIVE Hive is a Hadoop-based data warehousing–like framework originally developed by Facebook. It allows users to write queries in an SQL-like language called HiveQL, which are then converted to MapReduce. This allows SQL programmers with no MapReduce experience to use the warehouse and makes it easier to integrate with business intelligence (BI) and visualization tools such as Microstrategy, Tableau, Revolutions Analytics, and so forth.

PIG Pig is a Hadoop-based query language developed by Yahoo! It is relatively easy to learn and is adept at very deep, very long data pipelines (a limitation of SQL).

HBASE HBase is a nonrelational database that allows for low-latency, quick lookups in Hadoop. It adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts, and deletes. eBay and Facebook use HBase heavily.

FLUME Flume is a framework for populating Hadoop with data. Agents are populated throughout one's IT infrastructure—inside Web servers, application servers, and mobile devices, for example—to collect data and integrate it into Hadoop.

OOZIE Oozie is a workflow processing system that lets users define a series of jobs written in multiple languages—such as MapReduce, Pig, and Hive—and then intelligently link them to one another. Oozie allows users to specify, for example, that a particular query is only to be initiated after specified previous jobs on which it relies for data are completed.

AMBARI Ambari is a Web-based set of tools for deploying, administering, and monitoring Apache Hadoop clusters. Its development is being led by engineers from Hortonworks, which includes Ambari in its Hortonworks Data Platform.

AVRO Avro is a data serialization system that allows for encoding the schema of Hadoop files. It is adept at parsing data and performing removed procedure calls.

MAHOUT Mahout is a data mining library. It takes the most popular data mining algorithms for performing clustering, regression testing, and statistical modeling and implements them using the MapReduce model.

SQOOP Sqoop is a connectivity tool for moving data from non-Hadoop data stores—such as relational databases and data warehouses—into Hadoop. It allows users to specify the target location inside of Hadoop and instructs Sqoop to move data from Oracle, Teradata, or other relational databases to the target.

HCATALOG HCatalog is a centralized metadata management and sharing service for Apache Hadoop. It allows for a unified view of all data in Hadoop clusters and allows diverse tools, including Pig and Hive, to process any data elements without needing to know physically where in the cluster the data is stored.

Hadoop: The Pros and Cons

The main benefit of Hadoop is that it allows enterprises to process and analyze large volumes of unstructured and semistructured data, heretofore inaccessible to them, in a cost- and time-effective manner. Because Hadoop clusters can scale to petabytes and even exabytes of data, enterprises no longer must rely on sample data sets but can process and analyze *all* relevant data. Data scientists can apply an iterative approach to analysis, continually refining and testing queries to uncover previously unknown insights. It is also inexpensive to get started with Hadoop. Developers can download the Apache Hadoop distribution for free and begin experimenting with Hadoop in less than a day.

The downside to Hadoop and its myriad components is that they are immature and still developing. As with any young, raw technology, implementing and managing Hadoop clusters and performing advanced analytics on large volumes of unstructured data require significant expertise, skill, and training. Unfortunately, there is currently a dearth of Hadoop developers and data scientists available, making it impractical for many enterprises to maintain and take advantage of complex Hadoop clusters. Further, as Hadoop's myriad components are improved on by the community and new components are created, there is, as with any immature open source technology/approach, a risk of forking. Finally, Hadoop is a batch-oriented framework, meaning it does not support real-time data processing and analysis.

The good news is that some of the brightest minds in IT are contributing to the Apache Hadoop project, and a new generation of Hadoop developers and data scientists is coming of age. As a result, the technology is advancing rapidly, becoming both more powerful and easier to implement and manage. An ecosystem of vendors, both Hadoop-focused start-ups like Cloudera and Hortonworks and well-worn IT stalwarts like IBM, Microsoft, Teradata, and Oracle are working to offer commercial, enterprise-ready Hadoop distributions, tools, and services to make deploying and managing the technology a practical reality for the traditional enterprise. Other bleeding edge start-ups are working to perfect NoSQL (Not Only SQL) data stores capable of delivering near–real-time insights in conjunction with Hadoop. Technology Insights 9.2 provides a few facts to clarify some misconceptions about Hadoop.

TECHNOLOGY INSIGHTS 9.2 A Few Demystifying Facts about Hadoop

Although Hadoop and related technologies have been around for more than five years now, most people still have several misconceptions about Hadoop and related technologies such as MapReduce and Hive. The following list of 10 facts intends to clarify what Hadoop is and does relative to BI, as well as in which business and technology situations Hadoop-based BI, data warehousing, and analytics can be useful (Russom, 2013).

Fact #1. Hadoop consists of multiple products. We talk about Hadoop as if it's one monolithic software, whereas it is actually a family of open source products and technologies overseen by the Apache Software Foundation (ASF). (Some Hadoop products are also available via vendor distributions; more on that later.)

The Apache Hadoop library includes (in BI priority order) HDFS, MapReduce, Hive, Hbase, Pig, Zookeeper, Flume, Sqoop, Oozie, Hue, and so on. You can combine these in various ways, but HDFS and MapReduce (perhaps with Hbase and Hive) constitute a useful technology stack for applications in BI, data warehousing, and analytics.

Fact #2. Hadoop is open source but available from vendors, too. Apache Hadoop's open source software library is available from ASF at **apache.org**. For users desiring a more enterprise-ready package, a few vendors now offer Hadoop distributions that include additional administrative tools and technical support.

Fact #3. Hadoop is an ecosystem, not a single product. In addition to products from Apache, the extended Hadoop ecosystem includes a growing list of vendor products that integrate with or expand Hadoop technologies. One minute on your favorite search engine will reveal these.

Fact #4. HDFS is a file system, not a database management system (DBMS). Hadoop is primarily a distributed file system and lacks capabilities we would associate with a DBMS, such as indexing, random access to data, and support for SQL. That's okay, because HDFS does things DBMSs cannot do.

Fact #5. Hive resembles SQL but is not standard SQL. Many of us are handcuffed to SQL because we know it well and our tools demand it. People who know SQL can quickly learn to hand-code Hive, but that doesn't solve compatibility issues with SQL-based tools. TDWI feels that over time, Hadoop products will support standard SQL, so this issue will soon be moot.

Fact #6. Hadoop and MapReduce are related but don't require each other. Developers at Google developed MapReduce before HDFS existed, and some variations of MapReduce work with a variety of storage technologies, including HDFS, other file systems, and some DBMSs.

Fact #7. MapReduce provides control for analytics, not analytics per se. MapReduce is a general-purpose execution engine that handles the complexities of network communication, parallel programming, and fault tolerance for any kind of application that you can hand code—not just analytics.

Fact #8. Hadoop is about data diversity, not just data volume. Theoretically, HDFS can manage the storage and access of any data type as long as you can put the data in a file and copy that file into HDFS. As outrageously simplistic as that sounds, it's largely true, and it's exactly what brings many users to Apache HDFS.

Fact #9. Hadoop complements a DW; it's rarely a replacement. Most organizations have designed their DW for structured, relational data, which makes it difficult to wring BI value from unstructured and semistructured data. Hadoop promises to complement DWs by handling the multistructured data types most DWs can't.

Fact #10. Hadoop enables many types of analytics, not just Web analytics. Hadoop gets a lot of press about how Internet companies use it for analyzing Web logs and other Web data, but other use cases exist. For example, consider the Big Data coming from sensory devices, such as robotics in manufacturing, RFID in retail, or grid monitoring in utilities. Older analytic applications that need large data samples—such as customer-base segmentation, fraud detection, and risk analysis—can benefit from the additional Big Data managed by Hadoop. Likewise, Hadoop's additional data can expand 360-degree views to create a more complete and granular view.

NoSQL

A related new style of database called **NoSQL** (Not Only SQL) has emerged to, like Hadoop, process large volumes of multistructured data. However, whereas Hadoop is adept at supporting large-scale, batch-style historical analysis, NoSQL databases are aimed, for the most part (though there are some important exceptions), at serving up discrete data stored among large volumes of multistructured data to end-user and automated Big Data applications. This capability is sorely lacking from relational database technology, which simply can't maintain needed application performance levels at a Big Data scale.

In some cases, NoSQL and Hadoop work in conjunction. The aforementioned HBase, for example, is a popular NoSQL database modeled after Google BigTable that is often deployed on top of HDFS, the Hadoop Distributed File System, to provide low-latency, quick lookups in Hadoop. The downside of most NoSQL databases

today is that they trade ACID (atomicity, consistency, isolation, durability) compliance for performance and scalability. Many also lack mature management and monitoring tools. Both of these shortcomings are in the process of being overcome by the open source NoSQL communities and a handful of vendors that are attempting to commercialize the various NoSQL databases. NoSQL databases currently available include HBase, Cassandra, MongoDB, Accumulo, Riak, CouchDB, and DynamoDB, among others. Application Case 9.3 shows the use of NoSQL databases at eBay. Although the case is a few years old, we include it to give you a flavor of how multiple datasets come together. Application Case 9.4 illustrates a social media application where the Hadoop infrastructure was used to compile a corpus of messages on Twitter to understand which types of users engage in which type of support for healthcare patients seeking information about chronic mental diseases.

Application Case 9.3 eBay's Big Data Solution

eBay is one of the world's largest online marketplaces, enabling the buying and selling of practically anything. One of the keys to eBay's extraordinary success is its ability to turn the enormous volumes of data it generates into useful insights that its customers can glean directly from the pages they frequent. To accommodate eBay's explosive data growth—its data centers perform billions of reads and writes each day—and due to the increasing demand to process data at blistering speeds, eBay needed a solution that did not have the typical bottlenecks, scalability issues, and transactional constraints associated with common relational database approaches. The company also needed to perform rapid analysis on a broad assortment of the structured and unstructured data it captured.

The Solution: Integrated Real-Time Data and Analytics

Its Big Data requirements brought eBay to NoSQL technologies, specifically Apache Cassandra and DataStax Enterprise. Along with Cassandra and its high-velocity data capabilities, eBay was also drawn to the integrated Apache Hadoop analytics that come with DataStax Enterprise. The solution incorporates a scale-out architecture that enables eBay to deploy multiple DataStax Enterprise clusters across several different data centers using commodity hardware. The end result is that eBay is now able to more cost effectively process massive amounts of data at very high speeds, at very high velocities, and achieve far more than they were able to with the higher cost proprietary system they had been using. Currently, eBay is managing a sizable portion of its data center needs—250TBs+ of storage—in Apache Cassandra and DataStax Enterprise clusters.

Additional technical factors that played a role in eBay's decision to deploy DataStax Enterprise so widely include the solution's linear scalability, high availability with no single point of failure, and outstanding write performance.

Handling Diverse Use Cases

eBay employs DataStax Enterprise for many different use cases. The following examples illustrate some of the ways the company is able to meet its Big Data needs with the extremely fast data handling and analytics capabilities the solution provides. Naturally, eBay experiences huge amounts of write traffic, which the Cassandra implementation in DataStax Enterprise handles more efficiently than any other RDBMS or NoSQL solution. eBay currently sees 6 billion+ writes per day across multiple Cassandra clusters and 5 billion+ reads (mostly offline) per day as well.

One use case supported by DataStax Enterprise involves quantifying the social data eBay displays on its product pages. The Cassandra distribution in DataStax Enterprise stores all the

(*Continued*)

Application Case 9.3 (Continued)

information needed to provide counts for "like," "own," and "want" data on eBay product pages. It also provides the same data for the eBay "Your Favorites" page that contains all the items a user likes, owns, or wants, with Cassandra serving up the entire "Your Favorites" page. eBay provides this data through Cassandra's scalable counters feature.

Load balancing and application availability are important aspects to this particular use case. The DataStax Enterprise solution gave eBay architects the flexibility they needed to design a system that enables any user request to go to any data center, with each data center having a single DataStax Enterprise cluster spanning those centers. This design feature helps balance the incoming user load and eliminates any possible threat to application downtime. In addition to the line of business data powering the Web pages its customers visit, eBay is also able to perform high-speed analysis with the ability to maintain a separate data center running Hadoop nodes of the same DataStax Enterprise ring (see Figure 9.6).

Another use case involves the Hunch (an eBay sister company) "taste graph" for eBay users and items, which provides customer recommendations based on user interests. eBay's Web site is essentially a graph between all users and the items for sale. All events (bid, buy, sell, and list) are captured by eBay's systems and stored as a graph in Cassandra. The application sees more than 200 million writes daily and holds more than 40 billion pieces of data.

eBay also uses DataStax Enterprise for many time-series use cases in which processing high-volume, real-time data is a foremost priority. These include mobile notification logging and tracking (every time eBay sends a notification to a mobile phone or device it is logged in Cassandra), fraud detection, SOA request/response payload logging, and RedLaser (another eBay sister company) server logs and analytics.

Across all of these use cases is the common requirement of uptime. eBay is acutely aware of the need to keep their business up and open for business, and DataStax Enterprise plays a key part in that through its support of high-availability clusters. "We have to be ready for disaster recovery all the time. It's really great that Cassandra allows for active-active multiple data centers where we can read and write data anywhere, anytime," says eBay architect Jay Patel.

Questions for Discussion

1. Why did eBay need a Big Data solution?
2. What were the challenges, the proposed solution, and the obtained results?

Source: DataStax. Customer case studies. **datastax.com/resources/casestudies/eBay** (accessed October 2018).

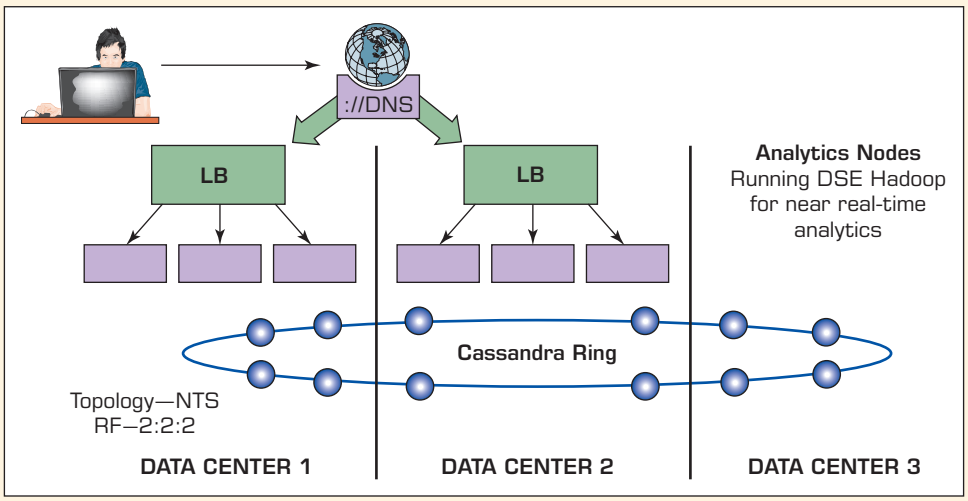


FIGURE 9.6 **eBay's Multi–Data Center Deployment.** *Source:* DataStax.

Application Case 9.4

Understanding Quality and Reliability of Healthcare Support Information on Twitter

On the Internet today, all users have the power to contribute as well as consume information. This power is used in many ways. On social network platforms such as Twitter, users are able to post information about their health condition as well as receive help on how best to manage those health conditions. Many users have wondered about the quality of information disseminated on social network platforms. Whereas the ability to author and disseminate health information on Twitter seems valuable to many users who use it to seek support for their disease, the authenticity of such information, especially when it originates from lay individuals, has been in doubt. Many users have asked, "How do I verify and trust information from nonexperts about how to manage a vital issue like my health condition?"

What types of users share and discuss what type of information? Do users with a large following discuss and share the same type of information as users with a smaller following? The number of followers of a user relate to the influence of a user. Characteristics of the information are measured in terms of quality and objectivity of the Tweet posted. A team of data scientists set out to explore the relationship between the number of followers a user had and the characteristics of information the user disseminated (Asamoah & Sharda, 2015).

Solution

Data was extracted from the Twitter platform using Twitter's API. The data scientists adapted the knowledge-discovery and data management model to manage and analyze this large set of data. The model was optimized for managing and analyzing Big Data derived from a social network platform and included phases for gaining domain knowledge, developing an appropriate Big Data platform, data acquisition and storage, data cleaning, data validation, data analysis, and results and deployment.

Technology Used

The tweets were extracted, managed, and analyzed using Cloudera's distribution of the Apache Hadoop. The Apache Hadoop framework has several subprojects that support different kinds of data management activities. For instance, the Apache Hive subproject supported the reading, writing, and managing of the large tweet data. Data analytics tools such as Gephi were used for social network analysis and R for predictive modeling. They conducted two parallel analyses; social network analysis to understand the influence network on the platform and text mining to understand the content of tweets posted by users.

What Was Found?

As noted earlier, tweets from both influential and noninfluential users were collected and analyzed. The results showed that the quality and objectivity of information disseminated by influential users was higher than that disseminated by noninfluential users. They also found that influential users controlled the flow of information in a network and that other users were more likely to follow their opinion on a subject. There was a clear difference between the type of information support provided by influential users versus the others. Influential users discussed more objective information regarding the disease management—things such as diagnoses, medications, and formal therapies. Noninfluential users provided more information about emotional support and alternative ways of coping with such diseases. Thus, a clear difference between influential users and the others was evident.

From the nonexperts' perspective, the data scientists portray how healthcare provision can be augmented by helping patients identify and use valuable resources on the Web for managing their disease condition. This work also helps identify how nonexperts can locate and filter healthcare information that may not necessarily be beneficial to the management of their health condition.

(*Continued*)

Application Case 9.4 (Continued)

Questions for Discussion

1. What was the data scientists' main concern regarding health information that is disseminated on the Twitter platform?
2. How did the data scientists ensure that nonexpert information disseminated on social media could indeed contain valuable health information?
3. Does it make sense that influential users would share more objective information whereas less influential users could focus more on subjective information? Why?

Sources: D. Asamoah & R. Sharda. (2015). "Adapting CRISP-DM Process for Social Network Analytics: Application to Healthcare." In *AMCIS 2015 Proceedings*. **aisel.aisnet.org/amcis2015/BizAnalytics/GeneralPresentations/33/** (accessed October 2018). Sarasohn-Kahn, J. (2008). *The Wisdom of Patients: Health Care Meets Online Social Media*. Oakland, CA: California HealthCare Foundation.

SECTION 9.4 REVIEW QUESTIONS

1. What are the common characteristics of emerging Big Data technologies?
2. What is MapReduce? What does it do? How does it do it?
3. What is Hadoop? How does it work?
4. What are the main Hadoop components? What functions do they perform?
5. What is NoSQL? How does it fit into the Big Data analytics picture?

9.5 BIG DATA AND DATA WAREHOUSING

There is no doubt that the emergence of Big Data has changed and will continue to change data warehousing in a significant way. Until recently, enterprise data warehouses (Chapter 3 and online supplements) were the centerpiece of all decision support technologies. Now, they have to share the spotlight with the newcomer, Big Data. The question that is popping up everywhere is whether Big Data and its enabling technologies such as Hadoop will replace data warehousing and its core technology RDBMS. Are we witnessing a data warehouse versus Big Data challenge (or from the technology standpoint, Hadoop versus RDBMS)? In this section we will explain why these questions have no basis—and at least justify that such an either-or choice is not a reflection of the reality at this point in time.

In the last decade or so, we have seen significant improvement in the area of computer-based decision support systems, which can largely be credited to data warehousing and technological advancements in both software and hardware to capture, store, and analyze data. As the size of the data increased, so did the capabilities of data warehouses. Some of these data warehousing advances included massively parallel processing (moving from one or few to many parallel processors), storage area networks (easily scalable storage solutions), solid-state storage, in-database processing, in-memory processing, and columnar (column-oriented) databases, just to name a few. These advancements helped keep the increasing size of data under control, while effectively serving analytics needs of the decision makers. What has changed the landscape in recent years is the variety and complexity of data, which made data warehouses incapable of keeping up. It is not the volume of the data but the variety and velocity that forced the world of IT to develop a new paradigm, which we now call "Big Data." Now that we have these two paradigms—data warehousing and Big Data—seemingly competing for the same job—turning data into actionable information—which one will prevail? Is this a fair question to ask? Or are we missing the big picture? In this section, we try to shed some light on this intriguing question.

As has been the case for many previous technology innovations, hype about Big Data and its enabling technologies like Hadoop and MapReduce is rampant. Nonpractitioners as well as practitioners are overwhelmed by diverse opinions. Yet others have begun to recognize that people are missing the point in claiming that Hadoop replaces relational databases and is becoming the new data warehouse. It is easy to see where these claims originate because both Hadoop and data warehouse systems can run in parallel, scale-up to enormous data volumes, and have shared-nothing architectures. At a conceptual level, one would think they are interchangeable. The reality is that they are not, and the differences between the two overwhelm the similarities. If they are not interchangeable, then how do we decide when to deploy Hadoop and when to use a data warehouse?

Use Cases for Hadoop

As we have covered earlier in this chapter, Hadoop is the result of new developments in computer and storage grid technologies. Using commodity hardware as a foundation, Hadoop provides a layer of software that spans the entire grid, turning it into a single system. Consequently, some major differentiators are obvious in this architecture:

Hadoop is the repository and refinery for raw data.
Hadoop is a powerful, economical, and active archive.

Thus, Hadoop sits at both ends of the large-scale data life cycle—first when raw data is born, and finally when data is retiring, but is still occasionally needed.

1. ***Hadoop as the repository and refinery.*** As volumes of Big Data arrive from sources such as sensors, machines, social media, and clickstream interactions, the first step is to capture all the data reliably and cost effectively. When data volumes are huge, the traditional single-server strategy does not work for long. Pouring the data into HDFS gives architects much needed flexibility. Not only can they capture hundreds of terabytes in a day, but they can also adjust the Hadoop configuration up or down to meet surges and lulls in data ingestion. This is accomplished at the lowest possible cost per gigabyte due to open source economics and leveraging commodity hardware.

   Because the data is stored on local storage instead of storage area networks, Hadoop data access is often much faster, and it does not clog the network with terabytes of data movement. Once the raw data is captured, Hadoop is used to refine it. Hadoop can act as a parallel "ETL engine on steroids," leveraging handwritten or commercial data transformation technologies. Many of these raw data transformations require the unraveling of complex freeform data into structured formats. This is particularly true with clickstreams (or Web logs) and complex sensor data formats. Consequently, a programmer needs to tease the wheat from the chaff, identifying the valuable signal in the noise.
2. ***Hadoop as the active archive.*** In a 2003 interview with ACM, Jim Gray claimed that hard disks could be treated as tape. Although it may take many more years for magnetic tape archives to be retired, today some portions of tape workloads are already being redirected to Hadoop clusters. This shift is occurring for two fundamental reasons. First, although it may appear inexpensive to store data on tape, the true cost comes with the difficulty of retrieval. Not only is the data stored offline, requiring hours if not days to restore, but tape cartridges themselves are also prone to degradation over time, making data loss a reality and forcing companies to factor in those costs. To make matters worse, tape formats change every couple of years, requiring organizations to either perform massive data migrations to the newest tape format or risk the inability to restore data from obsolete tapes.

Second, it has been shown that there is value in keeping historical data online and accessible. As in the clickstream example, keeping raw data on a spinning disk for a longer duration makes it easy for companies to revisit data when the context changes and new constraints need to be applied. Searching thousands of disks with Hadoop is dramatically faster and easier than spinning through hundreds of magnetic tapes. In addition, as disk densities continue to double every 18 months, it becomes economically feasible for organizations to hold many years' worth of raw or refined data in HDFS. Thus, the Hadoop storage grid is useful both in the preprocessing of raw data and the long-term storage of data. It's a true "active archive" because it not only stores and protects the data, but also enables users to quickly, easily, and perpetually derive value from it.

Use Cases for Data Warehousing

After nearly 30 years of investment, refinement, and growth, the list of features available in a data warehouse is quite staggering. Built on relational database technology using schemas and integrating BI tools, the major differences in this architecture are

Data warehouse performance
Integrated data that provides business value
Interactive BI tools for end users

1. ***Data warehouse performance.*** Basic indexing, found in open source databases, such as MySQL or Postgres, is a standard feature used to improve query response times or enforce constraints on data. More advanced forms such as materialized views, aggregate join indexes, cube indexes, and sparse join indexes enable numerous performance gains in data warehouses. However, the most important performance enhancement to date is the cost-based optimizer. The optimizer examines incoming SQL and considers multiple plans for executing each query as fast as possible. It achieves this by comparing the SQL request to the database design and extensive data statistics that help identify the best combination of execution steps. In essence, the optimizer is like having a genius programmer examine every query and tune it for the best performance. Lacking an optimizer or data demographic statistics, a query that could run in minutes may take hours, even with many indexes. For this reason, database vendors are constantly adding new index types, partitioning, statistics, and optimizer features. For the past 30 years, every software release has been a performance release. As we will note at the end of his section, Hadoop is now gaining on traditional data warehouses in terms of query performance.
2. ***Integrating data that provides business value.*** At the heart of any data warehouse is the promise to answer essential business questions. Integrated data is the unique foundation required to achieve this goal. Pulling data from multiple subject areas and numerous applications into one repository is the *raison d'être* for data warehouses. Data model designers and Extract, Transform, and Load (ETL) architects armed with metadata, data-cleansing tools, and patience must rationalize data formats, source systems, and semantic meaning of the data to make it understandable and trustworthy. This creates a common vocabulary within the corporation so that critical concepts such as "customer," "end of month," and "price elasticity" are uniformly measured and understood. Nowhere else in the entire IT data center is data collected, cleaned, and integrated as it is in the data warehouse.
3. ***Interactive BI tools.*** BI tools such as MicroStrategy, Tableau, IBM Cognos, and others provide business users with direct access to data warehouse insights. First, the business user can create reports and complex analysis quickly and easily using

these tools. As a result, there is a trend in many data warehouse sites toward end-user self-service. Business users can easily demand more reports than IT has staffing to provide. More important than self-service, however, is that the users become intimately familiar with the data. They can run a report, discover they missed a metric or filter, make an adjustment, and run their report again all within minutes. This process results in significant changes in business users' understanding of the business and their decision-making process. First, users stop asking trivial questions and start asking more complex strategic questions. Generally, the more complex and strategic the report, the more revenue and cost savings the user captures. This leads to some users becoming "power users" in a company. These individuals become wizards at teasing business value from the data and supplying valuable strategic information to the executive staff. Every data warehouse has anywhere from 2 to 20 power users. As noted in Section 9.8, all of these BI tools have begun to embrace Hadoop to be able to scale their offerings to larger data stores.

The Gray Areas (Any One of the Two Would Do the Job)

Even though there are several areas that differentiate one from the other, there are also gray areas where the data warehouse and Hadoop cannot be clearly discerned. In these areas either tool could be the right solution—either doing an equally good or a not-so-good job on the task at hand. Choosing one over the other depends on the requirements and the preferences of the organization. In many cases, Hadoop and the data warehouse work together in an information supply chain, and just as often, one tool is better for a specific workload (Awadallah & Graham, 2012). Table 9.1 illustrates the preferred platform (one versus the other, or equally likely) under a number of commonly observed requirements.

TABLE 9.1 When to Use Which Platform—Hadoop versus DW

Requirement	Data Warehouse	Hadoop
Low latency, interactive reports, and OLAP	☑	
ANSI 2003 SQL compliance is required	☑	☑
Preprocessing or exploration of raw unstructured data		☑
Online archives alternative to tape		☑
High-quality cleansed and consistent data	☑	☑
100s to 1,000s of concurrent users	☑	☑
Discover unknown relationships in the data		☑
Parallel complex process logic	☑	☑
CPU intense analysis	☑	
System, users, and data governance		☑
Many flexible programming languages running in parallel		☑
Unrestricted, ungoverned sandbox explorations		☑
Analysis of provisional data	☑	
Extensive security and regulatory compliance	☑	☑

Coexistence of Hadoop and Data Warehouse

There are several possible scenarios under which using a combination of Hadoop and relational DBMS-based data warehousing technologies makes more sense. Here are some of those scenarios (White, 2012):

1. ***Use Hadoop for storing and archiving multistructured data.*** A connector to a relational DBMS can then be used to extract required data from Hadoop for analysis by the relational DBMS. If the relational DBMS supports MapReduce functions, these functions can be used to do the extraction. The Vantage-Hadoop adaptor, for example, uses SQL-MapReduce functions to provide fast, two-way data loading between HDFS and the Vantage Database. Data loaded into the Vantage Database can then be analyzed using both SQL and MapReduce.
2. ***Use Hadoop for filtering, transforming, and/or consolidating multistructured data.*** A connector such as the Vantage-Hadoop adaptor can be used to extract the results from Hadoop processing to the relational DBMS for analysis.
3. ***Use Hadoop to analyze large volumes of multistructured data and publish the analytical results.*** In this application, Hadoop serves as the analytics platform, but the results can be posted back to the traditional data warehousing environment, a shared workgroup data store, or a common user interface.
4. ***Use a relational DBMS that provides MapReduce capabilities as an investigative computing platform.*** Data scientists can employ the relational DBMS (the Vantage Database system, for example) to analyze a combination of structured data and multistructured data (loaded from Hadoop) using a mixture of SQL processing and MapReduce analytic functions.
5. ***Use a front-end query tool to access and analyze data.*** Here, the data are stored in both Hadoop and the relational DBMS.

These scenarios support an environment where the Hadoop and relational DBMSs are separate from each other and connectivity software is used to exchange data between the two systems (see Figure 9.7). The direction of the industry over the next few years will likely be moving toward more tightly coupled Hadoop and relational DBMS-based data warehouse technologies—both software and hardware. Such integration provides

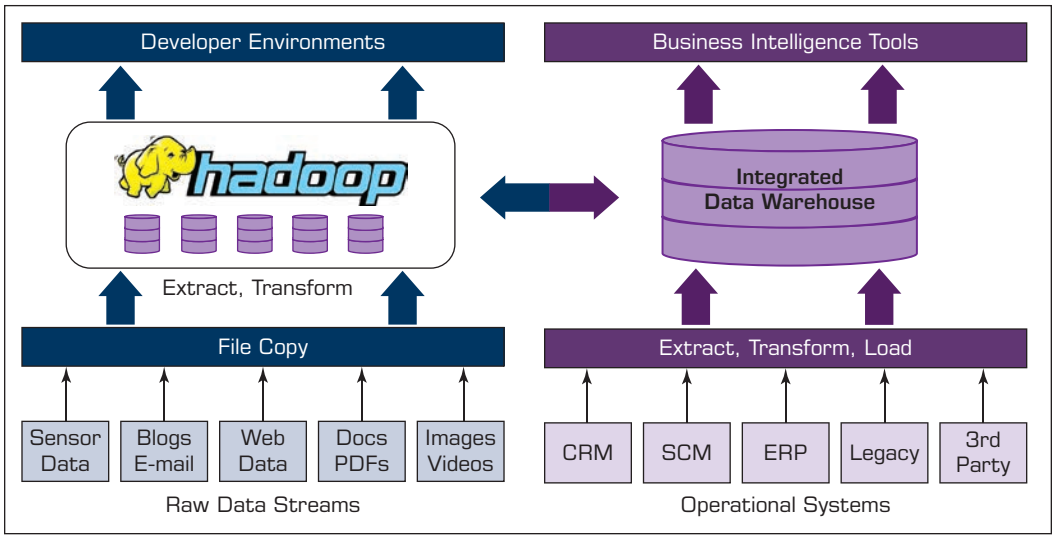


FIGURE 9.7 Coexistence of Hadoop and Data Warehouses. *Source:* "Hadoop and the Data Warehouse: When to Use Which, teradata, 2012." Used with permission from Teradata Corporation.

many benefits, including eliminating the need to install and maintain multiple systems, reducing data movement, providing a single metadata store for application development, and providing a single interface for both business users and analytical tools. The opening vignette (Section 9.1) provided an example of how data from a traditional data warehouse and two different unstructured data sets stored on Hadoop were integrated to create an analytics application to gain insight into a customer's interactions with a company before canceling an account. As a manager, you care about the insights you can derive from the data, not whether the data is stored in a structured data warehouse or a Hadoop cluster.

SECTION 9.5 REVIEW QUESTIONS

1. What are the challenges facing data warehousing and Big Data? Are we witnessing the end of the data warehousing era? Why or why not?
2. What are the use cases for Big Data and Hadoop?
3. What are the use cases for data warehousing and RDBMS?
4. In what scenarios can Hadoop and RDBMS coexist?

9.6 IN-MEMORY ANALYTICS AND APACHE SPARK™

Hadoop utilizes the batch processing framework and lacks real time processing capabilities. In the evolution of big data computing, in-memory analytics is an emerging processing technique to analyse data stored in in-memory databases. Because accessing data stored in memory is much faster than the data in hard disk, in-memory processing is more efficient than the batch processing. This also allows for the analytics of streaming data in real-time.

In-memory analytics have several applications where low latency execution is required. It can help build real-time dashboards for better insights and faster decision making. The real-time applications include understanding customer behaviour and engagement, forecasting stock price, optimizing airfare, predicting fraud, and several others.

The most popular tool supporting the in-memory processing is Apache Spark™. It is a unified analytics engine that can execute both batch and streaming data. Originally developed at University of California, Berkeley in 2009, Apache Spark™ uses in-memory computation to achieve high performance on large-scale data processing. By adopting an in-memory processing approach, Apache Spark™ runs faster than the traditional Apache Hadoop. Moreover, it can be interactively used from the Java, Scala, Python, R, and SQL shells for writing data management and machine learning applications. Apache Spark™ can run on Apache Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud. Besides, it can connect to different external data sources such as HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, and others.

Apache Spark™ can be used to create machine learning, fog computing, graph, streaming, and real-time analytics applications. Several big market players in the analytics sector have adopted Apache Spark™. Examples include Uber, Pinterest, Netflix, Yahoo, and eBay. Uber uses Apache Spark™ to detect fraudulent trips at scale. Pinterest measures user engagement in real-time using Apache Spark™. The recommendation engine of Netflix also utilizes the capabilities of Apache Spark™. Yahoo, one of the early adopters of Apache Spark™, has used it for creating business intelligence applications. Finally, eBay has used Apache Spark™ for data management and stream processing.

Application Case 9.5

Using Natural Language Processing to analyze customer feedback in TripAdvisor reviews

The TripAdvisor web platform contains information about hotels, restaurants, and other travel-related content. It also includes interactive travel forums that records the reviews of hotels or restaurants from the customers and managers. To improve the content on review forum, TripAdvisor decided to include tags to every travel attraction including restaurants and hotels. TripAdvisor collected reviews by sending out a review form to each of them, which basically had general review and some "yes" or "no" type questions. The yes/no responses from the customers resulted in different tags. Using the past information, the company decided to build a logistic regression model to forecast the yes/no response from a future customer and predict the tags. The problem is complex because each location has its own features. Using the past customers' experiences in the form of reviews, the textual information was used to train the model. The training model used the reviews on locations possessing the tag votes as well as the unlabeled reviews.

To create the model on the big data containing millions of reviews and hundreds of tags, the company adopted the Apache Spark™. Using the parallel processing and in-memory processing of Spark, a model was trained for each tag at each location. The data was partitioned by the location so as to minimize the communication among nodes. The entire process was implemented in an efficient manner.

Questions for Discussion

1. How did the predictive modelling help TripAdvisor?
2. Why was Spark used?

Compiled from: Palmucci, J., "Using Apache Spark for Massively Parallel NLP," at **http://engineering.tripadvisor.com/using-apache-spark-for-massively-parallel-nlp/** (accessed October 2018) and Dalininaa, R., "Using Natural Language Processing to Analyze Customer Feedback in Hotel Reviews," at **www.datascience.com/resources/notebooks/data-science-summarize-hotel-reviews** (accessed October 2018)

Architecture of Apache Spark™

Apache Spark™ works on a master-slave framework. There is a driver program that talks to the master node, also known as cluster manager, which manages the worker nodes. The execution of the tasks takes place in the worker nodes where executors run. The entry point of the engine is called a Spark Context. It acts as a bridge of communication between the application and the Spark execution environment as represented in Figure 9.8. As discussed earlier, Spark can run in different modes. In a standalone mode, it runs an application on different nodes in the cluster managed by the Spark itself. However, in a Hadoop mode, Spark uses Hadoop cluster to run jobs and leverage HDFS and MapReduce framework.

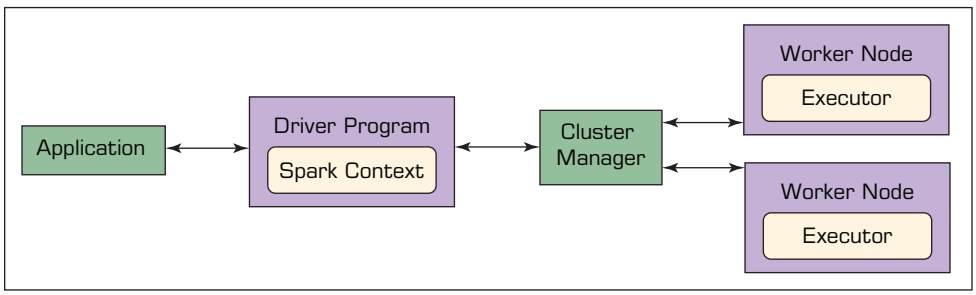


FIGURE 9.8 Apache Spark™ Architecture.

A very important component of **Apache Spark™** is Resilient Distributed Dataset, commonly known as RDD. It handles lineage, memory management, fault tolerance, and data partitioning across all nodes in a cluster. RDD provides several transformation functions like map, filter, and join that are performed on existing RDDs to create a new RDD. All transformations in Spark are lazy in nature, that is, Spark does not execute these operations until any action function is performed on data. The action functions (e.g., count, reduce) print or return value after an execution. This approach is called a **Lazy Evaluation**. In Spark Streaming, a series of RDDs, also known as a Dstream, are utilized to process streaming data.

Getting Started with Apache Spark™

In this section, we explain how to get started with Apache Spark on a Quick Start (QS) version of Cloudera Hadoop. It begins with downloading the latest version of Cloudera QS Virtual Machine (VM) and concludes with running your Spark query.

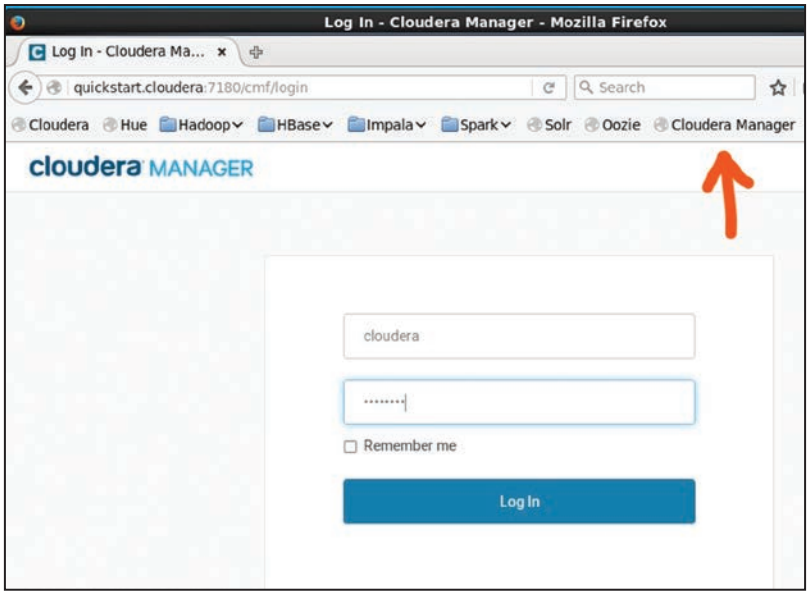
Hardware and Software Requirements Check

- A computer with 64-bit host Operating System (Windows or Linux) with at least 12 GB RAM for good performance
- VMware Workstation Player: Download and install the latest (free) version of the VMware Player from **www.vmware.com/products/workstation-player/workstation-player-evaluation.html**
- 8 GB memory for the VM | 20 GB free disk space
- 7-Zip: Extract (or unzip) the Cloudera Quick Start package using 7-Zip, available from: **www.7-zip.org/**

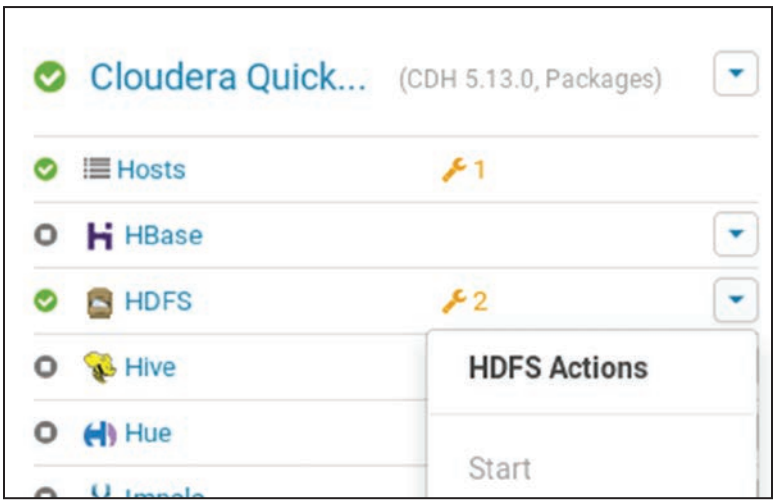
Steps to be followed to get started with Spark on Cloudera QS VM:

1. Download Cloudera QS VM from **www.cloudera.com/downloads/quickstart_vms/5-13.html**
2. Unzip it with 7-Zip. The downloaded file contains a VM machine.
3. Install VMware Workstation Player and turn it on. Now, open the Cloudera VM images through VMWare Player (Player > File > Open > full_path_of_vmx file).
4. Before turning on the VM, you must configure the memory and processer settings. The default memory on VM will be 4 GB RAM. Click on the "Edit virtual machine setting" to change the settings. Make sure the RAM is more than 8 GB and the number of processor cores is 2.
5. Turn on the machine. Cloudera has installed Hadoop and components on CentOS Linux.
6. A default user named "cloudera" with password "cloudera" is already available.
7. On the desktop of VM, open "Launch Cloudera Express." The engine will take a few minutes to get started.

8. Once started, open the web browser inside the VM. You will find an icon on the top of Cloudera Desktop.

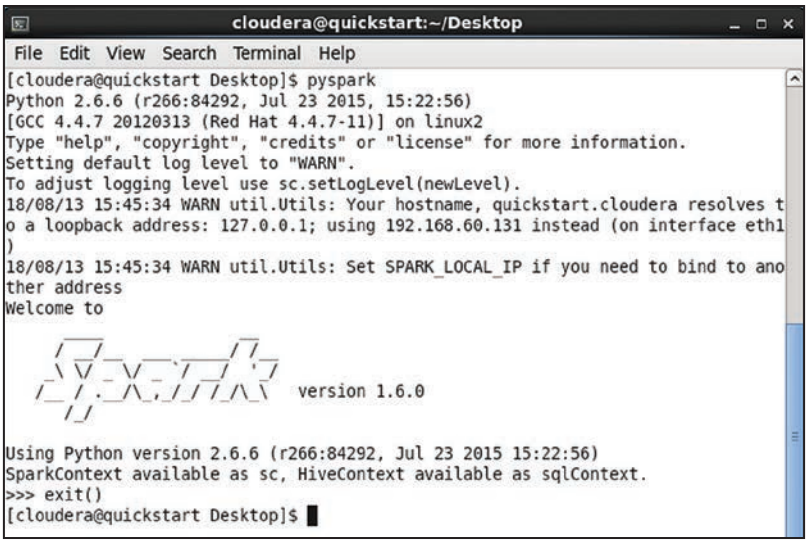


9. Login into Cloudera Manager using username "cloudera" and password "cloudera."
10. To use HDFS and map-reduce, we will want to start two services: HDFS and YARN using the drop-down menu in front of them.

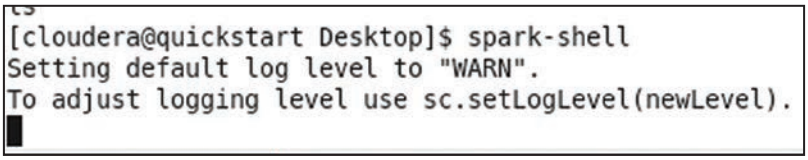


Courtesy of Mozilla Firefox.

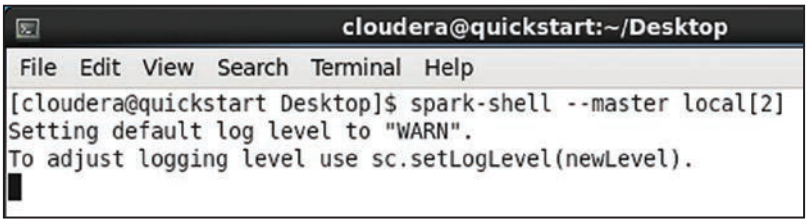
11. To turn on Spark, start the Spark service.
12. To run queries on Spark, we can use Python or Scala programming. Open Terminal by right clicking on the Desktop of VM.
13. Type *pyspark* to enter Python shell as shown in the screenshot below. To exit the Python shell, type *exit()*.



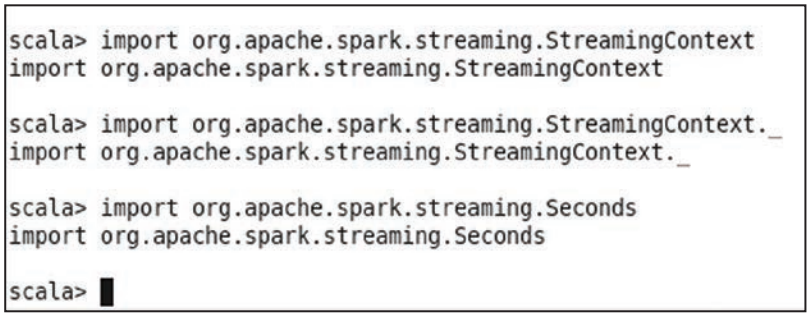
14. Type *spark-shell* to enter Scala Spark shell as shown in the screenshot below. To exit the Scala Spark shell, type *exit*.



15. From here onward, we describe steps to run a Spark streaming word count application where count of words will be calculated interactively. We run this application in Scala Spark shell. To use Spark Streaming interactively, we need to run Scala Spark shell with at least two threads. To do so, type *spark-shell –master local[2]* as shown in this screenshot.



a. Next, to run a streaming application, we need to import three related classes one by one as shown in this screenshot.



b. After importing the required classes, create a Spark Streaming Context *sss* with a batch duration of 10 seconds as in this screenshot.

```
scala> val sss = new StreamingContext(sc,Seconds(10))
sss: org.apache.spark.streaming.StreamingContext = org.apache.spark.streaming.Streami
ngContext@40ea1fe
```

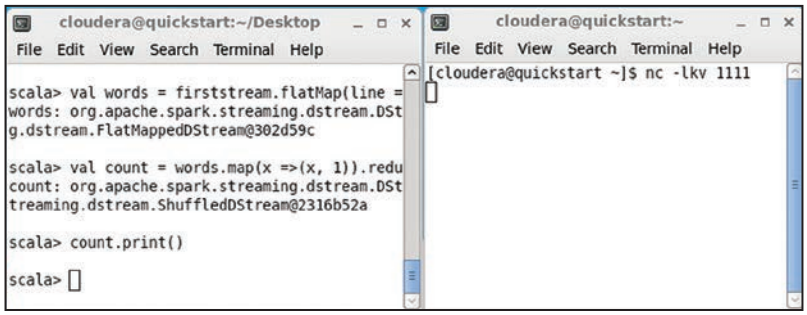
c. Create a discretized Stream (DStream), the basic abstraction in Spark Streaming, to read text from port 1111. It is presented in this screenshot.

```
scala> val firststream = sss.socketTextStream("localhost",1111)
firststream: org.apache.spark.streaming.dstream.ReceiverInputDStream[String] = org.ap
ache.spark.streaming.dstream.SocketInputDStream@7fa649a
```

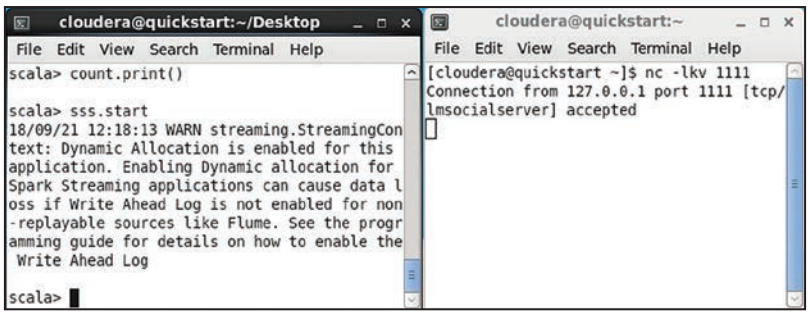
d. To count the occurrence of words on the stream, MapReduce codes shown in the screenshot are run. Then, *count.print()* command is used to print word count in the batch of 10 seconds.

```
scala> val words = firststream.flatMap(_.split(" "))
words: org.apache.spark.streaming.dstream.DStream[String] = org.apache.spark.str
eaming.dstream.FlatMappedDStream@7d9f5f12

scala> val pairs = words.map(word => (word, 1))
pairs: org.apache.spark.streaming.dstream.DStream[(String, Int)] = org.apache.sp
ark.streaming.dstream.MappedDStream@61614d4e

scala> val count = pairs.reduceByKey(_+_)
count: org.apache.spark.streaming.dstream.DStream[(String, Int)] = org.apache.sp
ark.streaming.dstream.ShuffledDStream@60519731

scala> count.print()
```

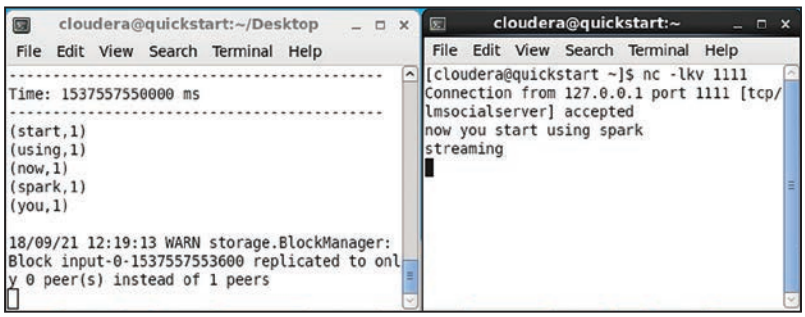
e. At this point, open a new terminal and run command ***nc − lkv* 1111** as shown in the right-hand-side terminal in this screenshot.



f. To start streaming context, run the *sss.start()* command in the Spark shell. This will result a connection of DStream *sss* with the socket (right-hand-side terminal).



g. In the final step, run *sss.awaitTermination()* in the Spark shell and start typing some words in the right-hand-side terminal as shown in this screenshot. After every 10 seconds, the word count pairs will be calculated in the Spark shell.



h. To stop the process, close the right-hand-side terminal and press CTRL + C in the left-hand-side terminal.

i. Because you may want to run the application again, all commands are listed here.

```
spark-shell –master local[2]
import org.apache.spark.streaming.StreamingContext
import org.apache.spark.streaming.StreamingContext._
import org.apache.spark.streaming.Seconds
val sss = new StreamingContext(sc,Seconds(10))
val firststream = sss.socketTextStream("localhost",1111)
val words = firststream.flatMap(_.split(" "))
val pairs = words.map(word => (word, 1))
val count = pairs.reduceByKey(_+_)
count.print()
sss.start()
sss.awaitTermination()
```

SECTION 9.6 REVIEW QUESTIONS

1. What are some of the unique features of Spark as compared to Hadoop?
2. Give examples of companies that have adopted Apache Spark. Find new examples online.
3. Run the exercise as described in this section. What do you learn from this exercise?

9.7 BIG DATA AND STREAM ANALYTICS

Along with volume and variety, as we have seen earlier in this chapter, one of the key characteristics that defines Big Data is velocity, which refers to the speed at which the data is created and streamed into the analytics environment. Organizations are looking for new means to process streaming data as it comes in to react quickly and accurately to problems and opportunities to please their customers and to gain a competitive advantage. In situations where data streams in rapidly and continuously, traditional analytics approaches that work with previously accumulated data (i.e., data at rest) often either arrive at the wrong decisions because of using too much out-of-context data, or they arrive at the correct decisions but too late to be of any use to the organization. Therefore, it is critical for a number of business situations to analyze the data soon after it is created and/or as soon as it is streamed into the analytics system.

The presumption that the vast majority of modern-day businesses are currently living by is that it is important and critical to record every piece of data because it might contain valuable information now or sometime in the near future. However, as long as

the number of data sources increases, the "store-everything" approach becomes harder and harder and, in some cases, not even feasible. In fact, despite technological advances, current total storage capacity lags far behind the digital information being generated in the world. Moreover, in the constantly changing business environment, real-time detection of meaningful changes in data as well as of complex pattern variations within a given short time window are essential to come up with the actions that better fit with the new environment. These facts become the main triggers for a paradigm that we call *stream analytics*. The stream analytics paradigm was born as an answer to these challenges, namely, the unbounded flows of data that cannot be permanently stored to be subsequently analyzed, in a timely and efficient manner, and complex pattern variations that need to be detected and acted on as soon as they happen.

Stream analytics (also called *data-in-motion analytics* and *real-time data analytics,* among other names) is a term commonly used for the analytic process of extracting actionable information from continuously flowing/streaming data. A stream is defined as a continuous sequence of data elements (Zikopoulos et al., 2013). The data elements in a stream are often called *tuples*. In a relational database sense, a tuple is similar to a row of data (a record, an object, an instance). However, in the context of semistructured or unstructured data, a tuple is an abstraction that represents a package of data, which can be characterized as a set of attributes for a given object. If a tuple by itself is not sufficiently informative for analysis or a correlation—or other collective relationships among tuples are needed—then a window of data that includes a set of tuples is used. A window of data is a finite number/sequence of tuples, where the windows are continuously updated as new data become available. The size of the window is determined based on the system being analyzed. Stream analytics is becoming increasingly more popular because of two things. First, time-to-action has become an ever-decreasing value, and second, we have the technological means to capture and process the data while it is created.

Some of the most impactful applications of stream analytics were developed in the energy industry, specifically for smart grid (electric power supply chain) systems. The new smart grids are capable of not only real-time creation and processing of multiple streams of data to determine optimal power distribution to fulfill real customer needs, but also generating accurate short-term predictions aimed at covering unexpected demand and renewable energy generation peaks. Figure 9.9 shows a depiction of a generic use case for streaming analytics in the energy industry (a typical smart grid application). The goal is to accurately predict electricity demand and production in real time by using streaming data that is coming from smart meters, production system sensors, and meteorological models. The ability to predict near future consumption/production trends and detect anomalies in real time can be used to optimize supply decisions (how much to produce, what sources of production to use, and optimally adjust production capacities) as well as to adjust smart meters to regulate consumption and favorable energy pricing.

Stream Analytics versus Perpetual Analytics

The terms *streaming* and *perpetual* probably sound like the same thing to most people, and in many cases they are used synonymously. However, in the context of intelligent systems, there is a difference (Jonas, 2007). Streaming analytics involves applying transaction-level logic to real-time observations. The rules applied to these observations take into account previous observations as long as they occurred in the prescribed window; these windows have some arbitrary size (e.g., last 5 seconds, last 10,000 observations). **Perpetual analytics**, on the other hand, evaluates every incoming observation against all prior observations, where there is no window size. Recognizing how the new observation relates to all prior observations enables the discovery of real-time insight.

Both streaming and perpetual analytics have their pros and cons and their respective places in the business analytics world. For example, sometimes transactional volumes

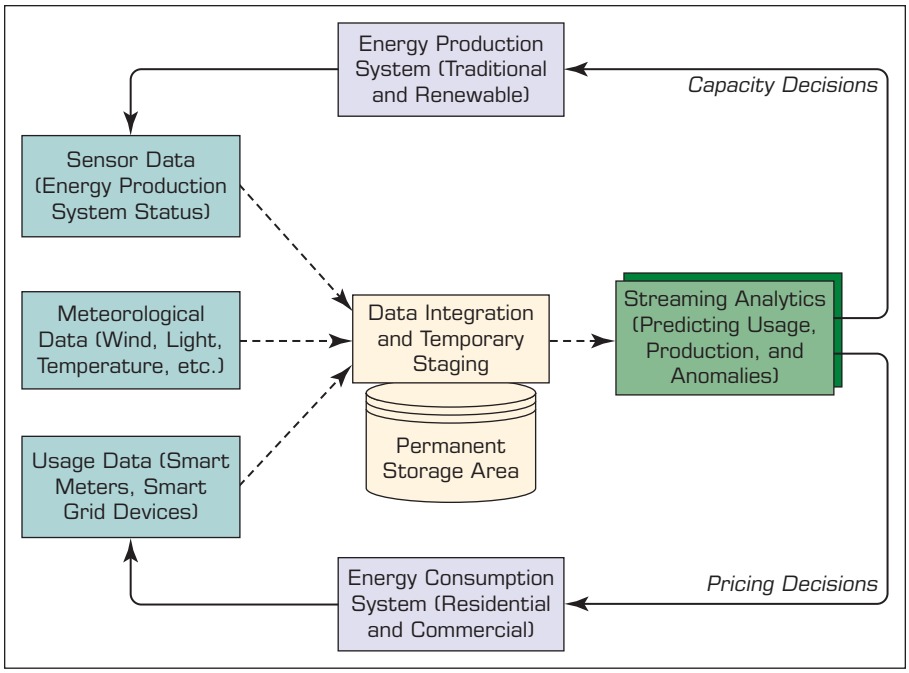


FIGURE 9.9 **A Use Case of Streaming Analytics in the Energy Industry.**

are high and the time-to-decision is too short, favoring nonpersistence and small window sizes, which translates into using streaming analytics. However, when the mission is critical and transaction volumes can be managed in real time, then perpetual analytics is a better answer. That way, one can answer questions such as "How does what I just learned relate to what I have known?" "Does this matter?" and "Who needs to know?"

Critical Event Processing

Critical event processing is a method of capturing, tracking, and analyzing streams of data to detect events (out of normal happenings) of certain types that are worthy of the effort. Complex event processing is an application of stream analytics that combines data from multiple sources to infer events or patterns of interest either before they actually occur or as soon as they happen. The goal is to take rapid actions to prevent (or mitigate the negative effects of) these events (e.g., fraud or network intrusion) from occurring, or in the case of a short window of opportunity, take full advantage of the situation within the allowed time (based on user behavior on an e-commerce site, create promotional offers that they are more likely to respond to).

These critical events may be happening across the various layers of an organization such as sales leads, orders, or customer service calls. Or, more broadly, they may be news items, text messages, social media posts, stock market feeds, traffic reports, weather conditions, or other kinds of anomalies that may have a significant impact on the well-being of the organization. An event may also be defined generically as a "change of state," which may be detected as a measurement exceeding a predefined threshold of time, temperature, or some other value. Even though there is no denying the value proposition of critical event processing, one has to be selective in what to measure, when to measure, and how often to measure. Because of the vast amount of information available about events, which is sometimes referred to as the *event cloud*, there is a possibility of overdoing it, in which case as opposed to helping the organization, it may hurt the operational effectiveness.

Data Stream Mining

Data stream mining, as an enabling technology for stream analytics, is the process of extracting novel patterns and knowledge structures from continuous, rapid data records. As we saw in the data mining chapter (Chapter 4), traditional data mining methods require the data to be collected and organized in a proper file format, and then processed in a recursive manner to learn the underlying patterns. In contrast, a data stream is a continuous flow of an ordered sequence of instances that in many applications of data stream mining can be read/processed only once or a small number of times using limited computing and storage capabilities. Examples of data streams include sensor data, computer network traffic, phone conversations, ATM transactions, Web searches, and financial data. Data stream mining is considered a subfield of data mining, machine learning, and knowledge discovery.

In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. Specialized machine-learning techniques (mostly derivative of traditional machine-learning techniques) can be used to learn this prediction task from labeled examples in an automated fashion. An example of such a prediction method was developed by Delen, Kletke, and Kim (2005), where they gradually built and refined a decision tree model by using a subset of the data at a time.

Applications of Stream Analytics

Because of its power to create insight instantly, helping decision makers to be on top of events as they unfold and allowing organizations to address issues before they become problems, the use of streaming analytics is on an exponentially increasing trend. The following are some of the application areas that have already benefited from stream analytics.

e-Commerce

Companies like Amazon and eBay (among many others) are trying to make the most out of the data that they collect while a customer is on their Web site. Every page visit, every product looked at, every search conducted, and every click made is recorded and analyzed to maximize the value gained from a user's visit. If done quickly, analysis of such a stream of data can turn browsers into buyers and buyers into shopaholics. When we visit an e-commerce Web site, even the ones where we are not a member, after a few clicks here and there we start to get very interesting product and bundle price offers. Behind the scenes, advanced analytics are crunching the real-time data coming from our clicks, and the clicks of thousands of others, to "understand" what it is that we are interested in (in some cases, even we do not know that) and make the most of that information by making creative offerings.

Telecommunications

The volume of data that come from call detail records (CDR) for telecommunications companies is astounding. Although this information has been used for billing purposes for quite some time now, there is a wealth of knowledge buried deep inside this Big Data that the telecommunications companies are just now realizing to tap. For instance, CDR data can be analyzed to prevent churn by identifying networks of callers, influencers, leaders, and followers within those networks and proactively acting on this information. As we all know, influencers and leaders have the effect of changing the perception of the followers within their network toward the service provider, either positively or negatively. Using social network analysis techniques, telecommunication companies are identifying the leaders and influencers and their network participants to better manage their customer base. In addition to churn analysis, such information can also be used to recruit new members and maximize the value of the existing members.

Application Case 9.6 Salesforce Is Using Streaming Data to Enhance Customer Value

Salesforce has expanded their Marketing Cloud services to include Predictive Scores and Predictive Audience features called the Marketing Cloud Predictive Journey. This addition uses real-time streaming data to enhance the customer engagement online. First, the customers are given a Predictive Score unique to them. This score is calculated from several different factors, including how long their browsing history is, if they clicked an e-mail link, if they made a purchase, how much they spent, how long ago did they make a purchase, or if they have ever responded to an e-mail or ad campaign. Once customers have a score, they are then segmented into different groups. These groups are given different marketing objectives and plans based on the predictive behaviors assigned to them. The scores and segments are updated and changed daily and give companies a better road map to target and achieve a desired response. These marketing solutions are more accurate and create more personalized ways companies can accommodate their customer retention methods.

Questions for Discussion

1. Are there areas in any industry where streaming data is irrelevant?
2. Besides customer retention, what are other benefits of using predictive analytics?

What Can We Learn from This Case?

Through the analysis of data acquired in the here and now, companies are able to make predictions and decisions about their consumers more rapidly. This ensures that businesses target, attract, and retain the right customers and maximize their value. Data acquired last week is not as beneficial as the data companies have today. Using relevant data makes our predictive analysis more accurate and efficient.

Sources: M. Amodio. (2015). "Salesforce Adds Predictive Analytics to Marketing Cloud. Cloud Contact Center." **www.cloudcontactcenterzone.com/topics/cloud-contact-center/articles/413611-salesforce-adds-predictive-analytics-marketing-cloud.htm** (accessed October 2018). J. Davis. (2015). "Salesforce Adds New Predictive Analytics to Marketing Cloud." *Information Week.* **www.informationweek.com/big-data/big-data-analytics/salesforce-adds-new-predictive-analytics-to-marketing-cloud/d/d-id/1323201** (accessed October 2018). D. Henschen. (2016). "Salesforce Reboots Wave Analytics, Preps IoT Cloud." *ZD Net.* **www.zdnet.com/article/salesforce-reboots-wave-analytics-preps-iot-cloud/** (accessed October 2018).

Continuous streams of data that come from CDR can be combined with social media data (sentiment analysis) to assess the effectiveness of marketing campaigns. Insight gained from these data streams can be used to rapidly react to adverse effects (which may lead to loss of customers) or boost the impact of positive effects (which may lead to maximizing purchases of existing customers and recruitment of new customers) observed in these campaigns. Furthermore, the process of gaining insight from CDR can be replicated for data networks using Internet protocol detail records. Because most telecommunications companies provide both of these service types, a holistic optimization of all offerings and marketing campaigns could lead to extraordinary market gains. Application Case 9.6 is an example of how **Salesforce.com** gets a better sense of its customers based upon an analysis of clickstreams.

Law Enforcement and Cybersecurity

Streams of Big Data provide excellent opportunities for improved crime prevention, law enforcement, and enhanced security. They offer unmatched potential when it comes to security applications that can be built in the space, such as real-time situational awareness, multimodal surveillance, cyber-security detection, legal wiretapping, video surveillance, and face recognition (Zikopoulos et al., 2013). As an application of

information assurance, enterprises can use streaming analytics to detect and prevent network intrusions, cyberattacks, and malicious activities by streaming and analyzing network logs and other Internet activity monitoring resources.

Power Industry

Because of the increasing use of smart meters, the amount of real-time data collected by power utilities is increasing exponentially. Moving from once a month to every 15 minutes (or more frequently), meter reading accumulates large quantities of invaluable data for power utilities. These smart meters and other sensors placed all around the power grid are sending information back to the control centers to be analyzed in real time. Such analyses help utility companies to optimize their supply chain decisions (e.g., capacity adjustments, distribution network options, real-time buying or selling) based on the up-to-the-minute consumer usage and demand patterns. In addition, utility companies can integrate weather and other natural conditions data into their analytics to optimize power generation from alternative sources (e.g., wind, solar) and to better forecast energy demand on different geographic granulations. Similar benefits also apply to other utilities such as water and natural gas.

Financial Services

Financial service companies are among the prime examples where analysis of Big Data streams can provide faster and better decisions, competitive advantage, and regulatory oversight. The ability to analyze fast-paced, high volumes of trading data at very low latency across markets and countries offers a tremendous advantage to making the split-second buy/sell decisions that potentially translate into big financial gains. In addition to optimal buy/sell decisions, stream analytics can also help financial service companies in real-time trade monitoring to detect fraud and other illegal activities.

Health Sciences

Modern-era medical devices (e.g., electrocardiograms and equipment that measure blood pressure, blood oxygen level, blood sugar level, and body temperature) are capable of producing invaluable streaming diagnostic/sensory data at a very fast rate. Harnessing this data and analyzing it in real time offers benefits—the kind that we often call "life and death"—unlike any other field. In addition to helping healthcare companies become more effective and efficient (and hence more competitive and profitable), stream analytics is also improving patient conditions and saving lives.

Many hospital systems all around the world are developing care infrastructures and health systems that are futuristic. These systems aim to take full advantage of what the technology has to offer, and more. Using hardware devices that generate high-resolution data at a very rapid rate, coupled with super-fast computers that can synergistically analyze multiple streams of data, increases the chances of keeping patients safe by quickly detecting anomalies. These systems are meant to help human decision makers make faster and better decisions by being exposed to a multitude of information as soon as it becomes available.

Government

Governments around the world are trying to find ways to be more efficient (via optimal use of limited resources) and effective (providing the services that people need and want). As the practices for e-government become mainstream, coupled with widespread use and access to social media, very large quantities of data (both structured and unstructured) are at the disposal of government agencies. Proper and timely use of these

Big Data streams differentiates proactive and highly efficient agencies from the ones who are still using traditional methods to react to situations as they unfold. Another way in which government agencies can leverage real-time analytics capabilities is to manage natural disasters such as snowstorms, hurricanes, tornadoes, and wildfires through a surveillance of streaming data coming from radar, sensors, and other smart detection devices. They can also use similar approaches to monitor water quality, air quality, and consumption patterns and detect anomalies before they become significant problems. Another area where government agencies use stream analytics is in traffic management in congested cities. By using the data coming from traffic flow cameras, GPS data coming from commercial vehicles, and traffic sensors embedded in roadways, agencies are able to change traffic light sequences and traffic flow lanes to ease the pain caused by traffic congestion problems.

SECTION 9.7 REVIEW QUESTIONS

1. What is a stream (in the Big Data world)?
2. What are the motivations for stream analytics?
3. What is stream analytics? How does it differ from regular analytics?
4. What is critical event processing? How does it relate to stream analytics?
5. Define *data stream mining*. What additional challenges are posed by data stream mining?
6. What are the most fruitful industries for stream analytics?
7. How can stream analytics be used in e-commerce?
8. In addition to what is listed in this section, can you think of other industries and/or application areas where stream analytics can be used?
9. Compared to regular analytics, do you think stream analytics will have more (or less) use cases in the era of Big Data analytics? Why?

9.8 BIG DATA VENDORS AND PLATFORMS

The Big Data vendor landscape is developing very rapidly. As is the case with many emerging technologies, even the terms change. Many Big Data technologies or solutions providers have rechristened themselves to be AI providers. In this section, we will do a quick overview of several categories of Big Data providers. Then we will briefly describe one provider's platform.

One way to study Big Data vendors and platforms is to go back to Chapter 1's analytics ecosystem (depicted in Figure 1.17). If we focus on some of the outermost petals of that analytics flower, we can see some categories of Big Data platform offerings. A more detailed classification of Big Data/AI providers is also included in Matt Turck's Big Data Ecosystem blog and the associated figure available at **http://mattturck.com/wp-content/uploads/2018/07/Matt_Turck_FirstMark_Big_Data_Landscape_2018_Final_reduced-768x539.png** (accessed October 2018). The reader is urged to check this site frequently to get updated versions of his view of the Big Data ecosystem.

In terms of technology providers, one thing is certain: everyone wants a bigger share of the technology spending pie and is thus willing to offer every single piece of technology, or partner with another provider so that the customer does not consider a competitor offering. Thus, many players seem to compete with each other by adding capabilities that their partners offer or by collaborating with their partners. In addition, there is always significant merger/acquisition activity. Finally, most vendors keep changing their products' names as the platforms evolve. This makes this specific section likely to be obsolete sooner than one might think. Recognizing all these caveats,

one highly aggregated way to group the Big Data providers is to use the following broad categories:

- Infrastructure Services Providers
- Analytics Solution Providers
- Legacy BI Providers Moving to Big Data

Infrastructure Services Providers

Big Data infrastructure was initially developed by two companies coming out of initial collaboration between Yahoo and Facebook. A number of vendors have developed their own Hadoop distributions, most based on the Apache open source distribution but with various levels of proprietary customization. Two market leaders were Cloudera (**cloudera.com**) and Hortonworks (**hortonworks.com**). Cloudera was started by Big Data experts including Hadoop creator Doug Cutting and former Facebook data scientist Jeff Hammerbacher. Hortonworks was spun out of Yahoo! These two companies have just (October 2018) announced a plan to merge into one company to provide a full suite of services in Big Data. The combined company will be able to offer Big Data services and be able to compete and partner with all other major providers. This makes it perhaps the largest independent provider of Hadoop distribution that provides on-premise Hadoop infrastructure, training, and support. MapR (**mapr.com**) offers its own Hadoop distribution that supplements HDFS with its proprietary network file system (NFS) for improved performance. Similarly, EMC was acquired by Dell to provide its own Big Data on-premise distribution. There are many other vendors that offer similar platforms with their own minor variations.

Another category of Hadoop distributors that add their own value-added services for customers are companies such as Datastax, Nutanix, VMWare, and so on. These companies deliver commercially supported versions of the various flavors of NoSQL. DataStax, for example, offers a commercial version of Cassandra that includes enterprise support and services, as well as integration with Hadoop and open source enterprise search. Many other companies provide Hadoop connectors and complementary tools aimed at making it easier for developers to move data around and within Hadoop clusters.

The next category of major infrastructure providers is the large cloud providers such as Amazon Web Services, Microsoft Azure, Google Cloud, and IBM Cloud. All of these companies offer storage and computing services but have invested heavily to provide Big Data and AI technology offerings. For example, Amazon AWS includes Hadoop and many other Big Data/AI capabilities (e.g., Amazon Neptune). Azure is a popular cloud provider for many analytics vendors, but Azure also offers its own Machine Learning and other capabilities. IBM and Google similarly offer their cloud services, but have major data science/AI offerings available, such as IBM Watson analytics and Google Tensor Flow, AutoML, and so on.

Analytics Solution Providers

The analytics layer of the Big Data stack is also experiencing significant development. Not surprisingly, all major traditional analytics and data service providers have incorporated Big Data analytics capabilities into their offerings. For example, Dell EMC, IBM Big Insights (now part of Watson), Microsoft Analytics, SAP's Hanna, Oracle Big Data, and Teradata have all integrated Hadoop, Streaming, IoT, and Spark capabilities into their platforms. IBM's BigInsights platform is based on Apache Hadoop, but includes numerous proprietary modules including the Netezza database, InfoSphere Warehouse, Cognos business intelligence tools, and SPSS data mining capabilities. It also offers IBM InfoSphere Streams, a platform designed for streaming Big Data analysis. With the success of Watson analytics brand, IBM has folded many of its analytics offerings in general and Big Data offerings in particular under the Watson label. Teradata Vantage similarly implements many of the commonly used analytics functions in the Big Data environment.

Further, as noted earlier, most of these platforms are also accessible through their own as well as public cloud providers. Rather than showing software details for all the platforms (which are quite similar anyway), we illustrate their description by using Teradata's newest offering, Teradata Vantage, in Technology Insights 9.3.

Business Intelligence Providers Incorporating Big Data

In this group we note several of the major BI software providers who, again not surprisingly, have incorporated Big Data technologies into their offerings. The major names to note in this space include SAS, Microstrategy, and their peers. For example, SAS Viya claims to perform in-memory analytics on massive data. Data visualization specialist Tableau Software has added Hadoop and Next Generation Data Warehouse connectivity to its product suite. Relatively newer players such as Qlik and Spotfire also are adapting their offerings to include Big data capabilities.

Application Case 9.7 illustrates an example of a Big data project where both IBM and Teradata analytics software capabilities were used in addition to pulling data from Google and Twitter.

Application Case 9.7 Using Social Media for Nowcasting Flu Activity

Infectious diseases impose a significant burden to the U.S. public health system. The rise of HIV/AIDS in the late 1970s, pandemic H1N1 flu in 2009, the H3N2 epidemic during the 2012–2013 winter season, the Ebola virus disease outbreak in 2015, and the Zika virus scare in 2016 have demonstrated the susceptibility of people to such contagious diseases. Virtually each year influenza outbreaks happen in various forms and result in consequences of varying impacts. The annual impact of seasonal influenza outbreaks in the United States is reported to be an average of 610,660 undiscounted life-years lost, 3.1 million hospitalized days, 31.4 million outpatient visits, and a total of $87.1 billion in economic burden. As a result of this growing trend, new data analytics techniques and technologies capable of detecting, tracking, mapping, and managing such diseases have come on the scene in recent years. In particular, digital surveillance systems have shown promise in their capacity to discover public health-seeking patterns and transform these discoveries into actionable strategies.

This project demonstrated that social media can be utilized as an effective method for early detection of influenza outbreaks. We used a Big Data platform to employ Twitter data to monitor influenza activity in the United States. Our Big Data analytics methods comprised temporal, spatial, and text mining. In the temporal analysis, we examined whether Twitter data could indeed be adapted for the nowcasting of influenza outbreaks. In spatial analysis, we mapped flu outbreaks to the geospatial property of Twitter data to identify influenza hotspots. Text analytics was performed to identify popular symptoms and treatments of flu that were mentioned in tweets.

The IBM InfoSphere BigInsights platform was employed to analyze two sets of flu activity data: Twitter data were used to monitor flu outbreaks in the United States, and Cerner HealthFacts data warehouse was used to track real-world clinical encounters. A huge volume of flu-related tweets was crawled from Twitter using Twitter Streaming API and was then ingested into a Hadoop cluster. Once the data were successfully imported, the JSON Query Language (JAQL) tool was used to manipulate and parse semistructured JavaScript Object Notation (JSON) data. Next, Hive was used to tabularize the text data and segregate the information for the spatial-temporal location analysis and visualization in R. The entire data mining process was implemented using MapReduce functions. We used the package BigR to submit the R scripts over the data stored in HDFS. The package BigR enabled us to benefit from the parallel computation of HDFS and to perform MapReduce operations. Google's Maps API libraries were used as a basic mapping tool to visualize the tweet locations.

(*Continued*)

Application Case 9.7 (Continued)

Our findings demonstrated that the integration of social media and medical records can be a valuable supplement to the existing surveillance systems. Our results confirmed that flu-related traffic on social media is closely related with the actual flu outbreak. This has been shown by other researchers as well (St Louis & Zorlu, 2012; Broniatowski, Paul, & Dredze, 2013). We performed a time-series analysis to obtain the spatial-temporal cross-correlation between the two trends (91%) and observed that clinical flu encounters lag behind online posts. In addition, our location analysis revealed several public locations from which a majority of tweets were originated. These findings can help health officials and governments to develop more accurate and timely forecasting models during outbreaks and to inform individuals about the locations that they should avoid during that time period.

Questions for Discussion

1. Why would social media be able to serve as an early predictor of flu outbreaks?
2. What other variables might help in predicting such outbreaks?
3. Why would this problem be a good problem to solve using Big Data technologies mentioned in this chapter?

Sources: A. H. Zadeh, H. M. Zolbanin, R. Sharda, & D. Delen. (2015). "Social Media for Nowcasting the Flu Activity: Spatial-Temporal and Text Analysis." *Business Analytics Congress, Pre-ICIS Conference*, Fort Worth, TX. D. A. Broniatowski, M. J. Paul, & M. Dredze. (2013). "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012–2013 Influenza Epidemic." *PloS One, 8*(12), e83672. P. A. Moran. (1950). "Notes on Continuous Stochastic Phenomena." *Biometrika*, 17–23.

TECHNOLOGY INSIGHTS 9.3 An Illustrative Big Data Technology Platform: Teradata Vantage™

Introduction

This description is adapted from content provided by Teradata, especially Sri Raghavan. Teradata Vantage is an advanced analytics platform embedded with analytic engines and functions, which can be implemented with preferred data science languages (e.g., SQL, Python, R) and tools (e.g., Teradata Studio, Teradata AppCenter, R Studio, Jupyter Notebook) on any data volume of any type by diverse analytics personas (e.g., Data Scientist, Citizen Data Scientist, Business Analyst) across multiple environments (On-Premises, Private Cloud, Public Cloud Marketplaces). There are five important conceptual pieces central to understanding Vantage: Analytics Engines and Functions, Data Storage and Access, Analytic Languages and Tools, Deployment, and Usage. Figure 9.10 illustrates the general architecture of Vantage and its interrelationships with other tools.

Analytic Engines and Functions

An analytic engine is a comprehensive framework that includes all the software components that are well integrated into a container (e.g., Docker) to deliver advanced analytics functionality that can be implemented by a well-defined set of user personas. An analytic engine's components include:

- Advanced Analytics functions
- Access points to data storage that can ingest multiple data types
- Integration into visualization and analytic workflow tools
- Built in management and monitoring tools
- Highly scalable and performant environment with established thresholds

It is advantageous to have an analytic engine as it delivers a containerized compute environment that can be separated from data storage. Furthermore, analytic engines can be tailored for access and use by specific personas (e.g., DS, Business Analyst).

There are three analytic engines in the first release of Vantage. These are NewSQL Engine, Machine Learning Engine, and Graph Engine.

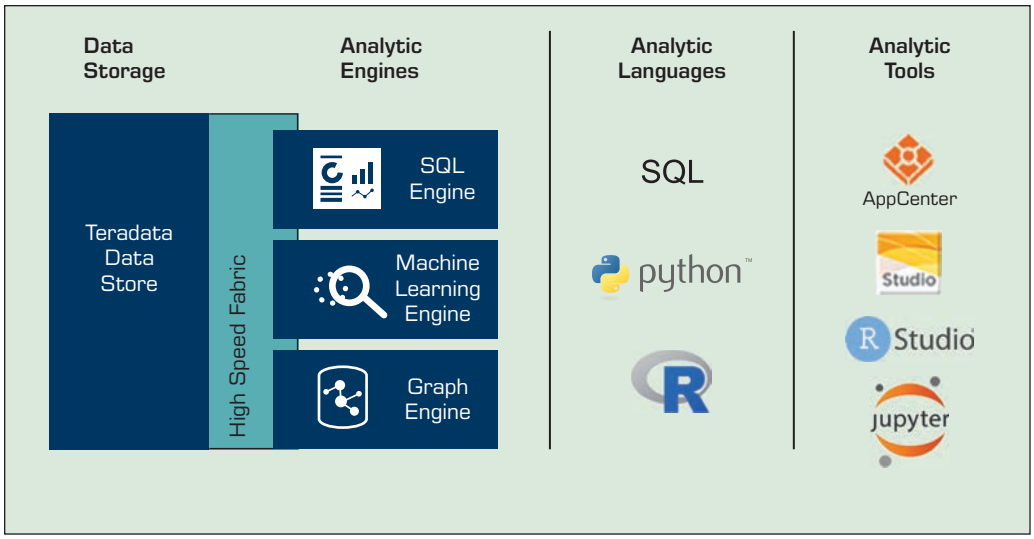


FIGURE 9.10 **Teradata Vantage Architecture.** *Source:* Teradata Corp.

The NewSQL engine includes embedded analytic functions. Teradata will continue to add more functions for the high-speed analytics processing required to operationalize analytics. New functions within the NewSQL engine include:

- nPath
- Sessionization
- Attribution
- Time series
- 4D analytics
- Scoring functions (e.g., Naïve Bayes, GLM, Decision Forests)

The Machine Learning engine delivers more than 120 prebuilt analytic functions for path, pattern, statistical, and text analytics to solve a range of business problems. Functions range from understanding sentiment to predictive part failure analysis.

The Graph engine provides a set of functions that discover relationships between people, products, and processes within a network. Graph analytics solve complex problems such as social network connections, influencer relationships, fraud detection, and threat identification.

Vantage embeds analytic engines close to the data, which eliminates the need to move data, allowing users to run their analytics against larger data sets without sampling and execute models with greater speed and frequency. This is made possible through the use of containers managed by Kubernetes, which allow businesses to easily manage and deploy new cutting-edge analytic engines, such as Spark and TensorFlow, both of which will be available in the near future. Another benefit of containers is the ability to scale out the engines.

From a user's perspective, Vantage is a unified analytic and data framework. Under the covers, it contains a cross-engine orchestration layer that pipelines the right data and analytic request to the right analytic engine across a high-speed data fabric. This enables a business analyst or data scientist, for example, to invoke analytic functions from different engines in a single application, such as Jupyter Notebook, without enduring the trouble of hopping from one analytic server or application to another. The result is a tightly integrated analytic implementation that's not restrained by functional or data silos.

Data Storage and Access: Teradata Vantage comes with a natively embedded Teradata MPP Database. Furthermore, a high-speed data fabric (Teradata QueryGrid™ and Presto™) connects the platform to external data sources that include third-party enterprise data warehouses (e.g., Oracle), open source data platforms (e.g., Hadoop), no-SQL databases

(e.g., Cassandra), and others. Data support ranges from relational, spatial, and temporal to XML, JSON, Avro, and time-series formats.

Analytic Languages and Tools: Teradata Vantage was built out of the recognition that analytics professionals such as Data Scientists and Business Analysts require a diverse set of languages and tools to process large data volumes to deliver analytic insights. Vantage includes languages such as SQL, R, and Python on which analytics functions can be executed through Teradata Studio, R Studio, and Jupyter Notebooks.

Deployment: Vantage platform provides the same analytic processing across deployment options, including the Teradata Cloud and public cloud, as well as on-premises installations on Teradata hardware or commodity hardware. It is also available as a service.

Usage: Teradata Vantage is intended to be used by multiple analytic personas. The ease of SQL ensures that citizen data scientists and business analysts can implement prebuilt analytic functions integrated into the analytic engines. The ability to invoke Teradata-supported packages such as dplyr and teradataml ensures that Data Scientists familiar with R and Python can execute analytic packages through R Studio and Jupyter notebooks, respectively, on the platform. Users who are not proficient at executing programs can invoke analytic functions codified into Apps built into Teradata AppCenter, an app building framework available in Vantage, to deliver compelling visualizations such as Sankey, Tree, Sigma diagrams, or word clouds.

Example Usage: A global retailer had a website that suboptimally delivered search results to potential buyers. With online purchases accounting for 25% of total sales, inaccurate search results negatively impacted the customer experience and the bottom line.

The retailer implemented Teradata machine learning algorithms, available in Teradata Vantage, to accumulate, parse, and classify search terms and phrases. The algorithms delivered the answers needed to identify search results that closely matched online customer needs. This led to more than $1.3 million in incremental revenue from high-value customers, as measured by purchase volumes, over a two-month holiday period.

Application Case 9.8 illustrates another application of Teradata Vantage where its advanced network analytics capabilities were deployed to analyze data from a large electronic medical records data warehouse.

Application Case 9.8 Analyzing Disease Patterns from an Electronic Medical Records Data Warehouse

The Center for Health Systems Innovation at Oklahoma State University has been given a massive data warehouse by Cerner Corporation, a major electronic medical records (EMRs) provider, to help develop analytic applications. The data warehouse contains EMRs on the visits of more than 50 million unique patients across U.S. hospitals (2000–2015). It is the largest and the industry's only relational database that includes comprehensive records with pharmacy, laboratory, clinical events, admissions, and billing data. The database also includes more than 2.4 billion laboratory results and more than 295 million orders for nearly 4,500 drugs by name and brand. It is one of the largest compilations of de-identified, real-world, HIPAA-compliant data of its type.

The EMRs can be used to develop multiple analytics applications. One application is to understand the relationships between diseases based on the information about the simultaneous diseases developed in the patients. When multiple diseases are present in a patient, the condition is called comorbidity. The comorbidities can be different across population groups. In an application (Kalgotra, Sharda, & Croff, 2017), the authors studied health disparities in terms of comorbidities by gender.

To compare the comorbidities, a network analysis approach was applied. A network is comprised of a defined set of items called nodes, which are linked to each other through edges. An edge represents a defined relationship between the

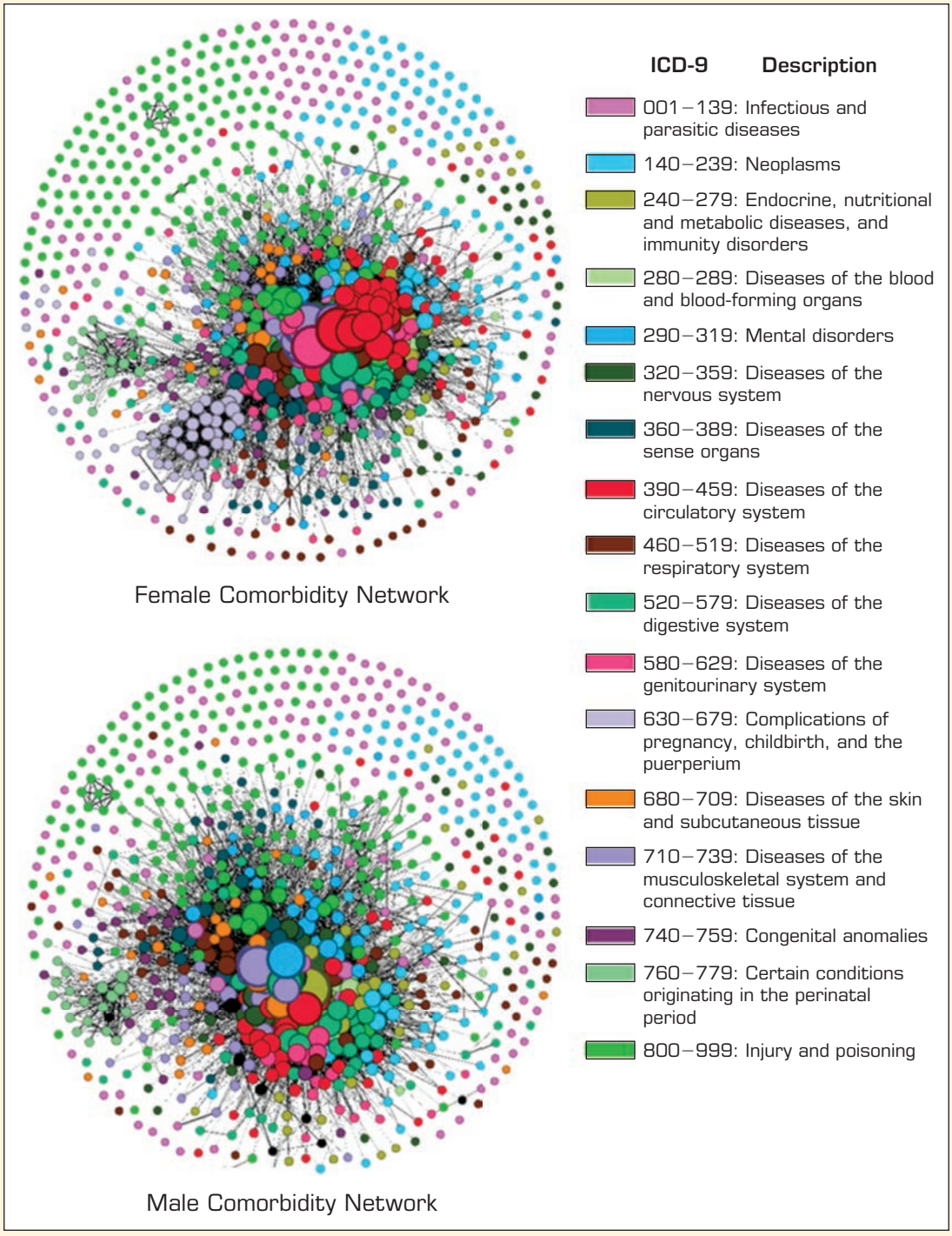


FIGURE 9.11 **Female and Male Comorbidity Networks.**

(*Continued*)

Application Case 9.8 (Continued)

nodes. A very common example of network is a friendship network in which individuals are connected to each other if they are friends. Other common networks are computer networks, Web page networks, road networks, and airport networks. To compare the comorbidities, networks of the diagnoses developed by men and women were created. The information about the diseases developed by each patient in the lifetime history was used to create a comorbidity network. For the analysis, 12 million female patients and 9.9 million male patients were used. To manage such a huge data set, Teradata Aster Big Data platform was used. To extract and prepare the network data, SQL, SQL-MR, and SQL-GR frameworks supported by Aster were utilized. To visualize the networks, Aster AppCenter and Gephi were used.

Figure 9.11 presents the female and male comorbidity networks. In these networks, nodes represent different diseases classified as the *International Classification of Diseases*, Ninth Revision, Clinical Modification (ICD-9-CM), aggregated at the three-digit level. Two diseases are linked based on the similarity calculated using Salton Cosine Index. The larger the size of a node, the greater the comorbidity of that disease. The female comorbidity network is denser than the male network. The number of nodes and edges in the female network are 899 and 14,810, respectively, whereas the number of nodes and edges in the male network are 839 and 12,498, respectively. The visualizations present a difference between the pattern of diseases developed in male and female patients. Specifically, females have more comorbidities of mental disorders than males. On the other hand, the strength of some disease associations between lipid metabolism and chronic heart disorders is stronger in males than females. Such health disparities present questions for biological, behavioural, clinical, and policy research.

The traditional database systems would be taxed in efficiently processing such a huge data set. The Teradata Aster made the analysis of data containing information on millions of records fairly fast and easy. Network analysis is often suggested as one method to analyze big data sets. It helps understand the data in one picture. In this application, the comorbidity network explains the relationship between diseases at one place.

Questions for Discussion

1. What could be the reasons behind the health disparities across gender?
2. What are the main components of a network?
3. What type of analytics was applied in this application?

Source: Kalgotra, P., Sharda, R., & Croff, J. M. (2017). Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *International Journal of Medical Informatics, 108*, 22–28.

As noted earlier, our goal in this section is to highlight some of the players in Big data technology space. In addition to the vendors listed above, there are hundreds of others in the categories identified earlier as well as very specific industry applications. Rather than listing these names here, we urge you to check the latest version of the Big Data analytics ecosystem at **http://mattturck.com/bigdata2018/** (accessed October 2018). Matt Turck's updated ecosystem diagram identifies companies in each cluster.

SECTION 9.8 REVIEW QUESTIONS

1. Identify some of the key Big Data technology vendors whose key focus is on-premise Hadoop platforms.
2. What is special about the Big Data vendor landscape? Who are the big players?
3. Search and identify the key similarity and differences between cloud providers' analytics offerings and analytics providers' presence on specific cloud platforms.
4. What are some of the features of a platform such as Teradata Vantage?

9.9 CLOUD COMPUTING AND BUSINESS ANALYTICS

Another emerging technology trend that business analytics users should be aware of is cloud computing. The National Institute of Standards and Technology (NIST) defines **cloud computing** as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction." Wikipedia (n.d., Cloud Computing) defines cloud computing as "a style of computing in which dynamically scalable and often virtualized resources are provided over the Internet. Users need not have knowledge of, experience in, or control over the technology infrastructures in the cloud that supports them." This definition is broad and comprehensive. In some ways, cloud computing is a new name for many previous, related trends: utility computing, application service provider grid computing, on-demand computing, software as a service (SaaS), and even older, centralized computing with dumb terminals. But the term *cloud computing* originates from a reference to the Internet as a "cloud" and represents an evolution of all of the previously shared/centralized computing trends. The Wikipedia entry also recognizes that cloud computing is a combination of several IT components as services. For example, *infrastructure as a service* (IaaS) refers to providing computing *platforms as a service* (PaaS), as well as all of the basic platform provisioning, such as management administration, security, and so on. It also includes SaaS, which includes applications to be delivered through a Web browser, whereas the data and the application programs are on some other server.

Although we do not typically look at Web-based e-mail as an example of cloud computing, it can be considered a basic cloud application. Typically, the e-mail application stores the data (e-mail messages) and the software (e-mail programs that let us process and manage e-mails). The e-mail provider also supplies the hardware/software and all of the basic infrastructure. As long as the Internet is available, one can access the e-mail application from anywhere in the cloud. When the application is updated by the e-mail provider (e.g., when Gmail updates its e-mail application), it becomes available to all customers. Social networking Web sites like Facebook, Twitter, and LinkedIn, are also examples of cloud computing. Thus, any Web-based general application is in a way an example of a cloud application. Another example of a general cloud application is Google Docs and Spreadsheets. This application allows a user to create text documents or spreadsheets that are stored on Google's servers and are available to the users anywhere they have access to the Internet. Again, no programs need to be installed as "the application is in the cloud." The storage space is also "in the cloud." Even Microsoft's popular office applications are all available in the cloud, with the user not needing to download any software.

A good general business example of cloud computing is **Amazon.com**'s Web services. **Amazon.com** has developed an impressive technology infrastructure for e-commerce as well as for BI, customer relationship management, and supply-chain management. It has built major data centers to manage its own operations. However, through **Amazon.com**'s cloud services, many other companies can employ these very same facilities to gain advantages of these technologies without having to make a similar investment. Like other cloud-computing services, a user can subscribe to any of the facilities on a pay-as-you-go basis. This model of letting someone else own the hardware and software but making use of the facilities on a pay-per-use basis is the cornerstone of cloud computing. A number of companies offer cloud-computing services, including **Salesforce.com**, IBM Cloud, Microsoft Azure, Google, Adobe, and many others.

Cloud computing, like many other IT trends, has resulted in new offerings in analytics. These options permit an organization to scale up its data warehouse and pay only for what it uses. The end user of a cloud-based analytics service may use one organization for analysis applications that, in turn, uses another firm for the platform or

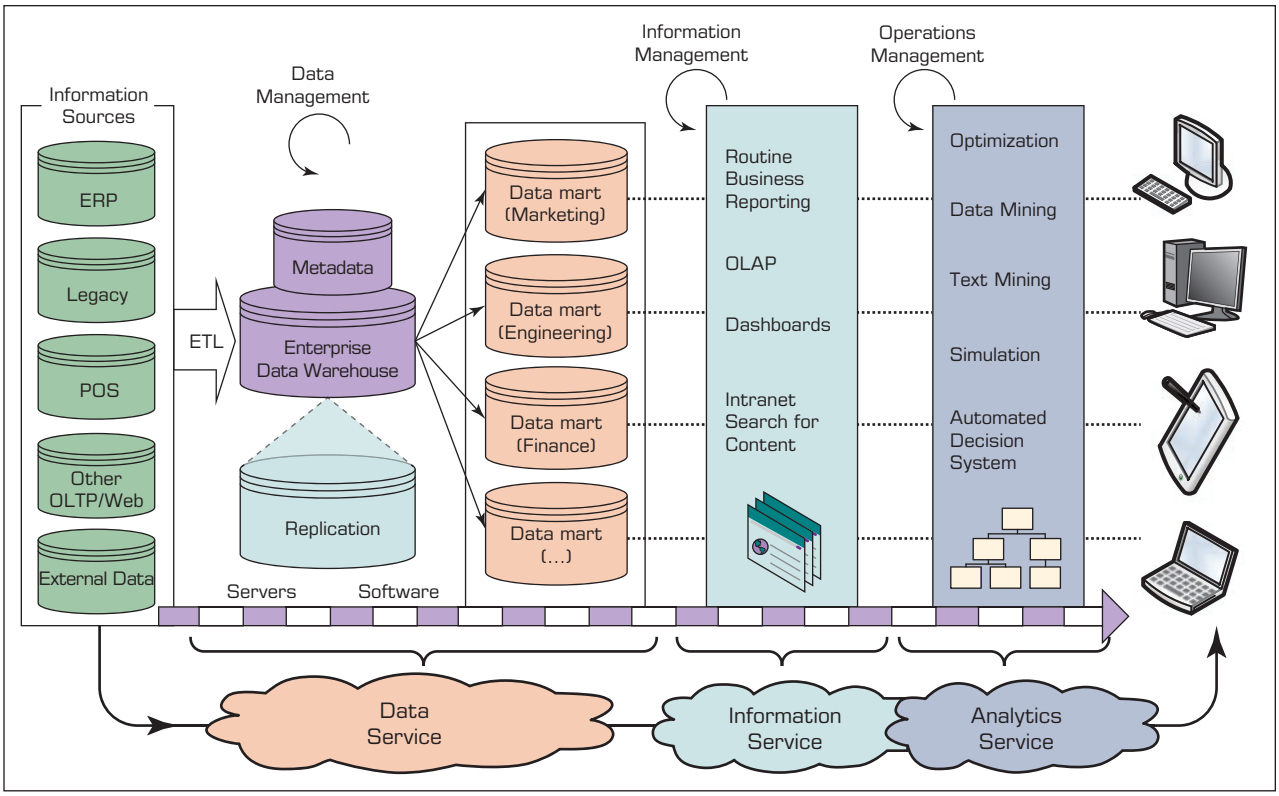


FIGURE 9.12 Conceptual Architecture of a Cloud-Oriented Support System. *Source:* Based on Demirkan, H., & Delen, D. (2013, April). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and Big Data in cloud. *Decision Support Systems, 55*(1), 412–421.

infrastructure. The next several paragraphs summarize the latest trends in the interface of cloud computing and BI/business analytics. A few of these statements are adapted from an early paper written by Haluk Demirkan and one of the coauthors of this book (Demirkan & Delen, 2013).

Figure 9.12 illustrates a conceptual architecture of a service-oriented decision support environment, that is, a cloud-based analytics system. This figure superimposes the cloud-based services on the general analytics architecture presented in previous chapters.

In service-oriented decision support solutions, (1) operational systems, (2) data warehouses, (3) online analytic processing, and (4) end-user components can be obtained individually or bundled and provided to the users as service. Any or all of these services can be obtained through the cloud. Because the field of cloud computing is fast evolving and growing at a rapid pace, there is much confusion about the terminology being used by various vendors and users. The labels vary from Infrastructure, Platform, Software, Data, Information, and Analytics as a Service. In the following, we define these services. Then we summarize the current technology platforms and highlight applications of each through application cases.

Data as a Service (DaaS)

The concept of data as a service basically advocates the view that "where data lives"—the actual platform on which the data resides—doesn't matter. Data can reside in a local computer or in a server at a server farm inside a cloud-computing environment.

With DaaS, any business process can access data wherever it resides. Data as a service began with the notion that data quality could happen in a centralized place, cleansing and enriching data and offering it to different systems, applications, or users, irrespective of where they were in the organization, computers, or on the network. This has now been replaced with master data management and customer data integration solutions, where the record of the customer (or product, or asset, etc.) may reside anywhere and is available as a service to any application that has the services allowing access to it. By applying a standard set of transformations to the various sources of data (for example, ensuring that gender fields containing different notation styles [e.g., M/F, Mr./Ms.] are all translated into male/female) and then enabling applications to access the data via open standards such as SQL, XQuery, and XML, service requestors can access the data regardless of vendor or system.

With DaaS, customers can move quickly thanks to the simplicity of the data access and the fact that they don't need extensive knowledge of the underlying data. If customers require a slightly different data structure or have location-specific requirements, the implementation is easy because the changes are minimal (agility). Second, providers can build the base with the data experts and outsource the analysis or presentation layers (which allows for very cost-effective user interfaces and makes change requests at the presentation layer much more feasible), and access to the data is controlled through the data services. It tends to improve data quality because there is a single point for updates.

Software as a Service (SaaS)

This model allows consumers to use applications and software that run on distant computers in the cloud infrastructure. Consumers need not worry about managing underlying cloud infrastructure and have to pay for the use of software only. All we need is a Web browser or an app on a mobile device to connect to the cloud. Gmail is an example of SaaS.

Platform as a Service (PaaS)

Using this model, companies can deploy their software and applications in the cloud so that their customers can use them. Companies don't have to manage resources needed to manage their applications in cloud-like networks, servers, storage, or operating systems. This reduces the cost of maintaining underlying infrastructure for running their software and also saves time for setting up this infrastructure. Now, users can focus on their business rather than focusing on managing infrastructure for running their software. Examples of PaaS are Microsoft Azure, Amazon EC2, and Google App Engine.

Infrastructure as a Service (IaaS)

In this model, infrastructure resources like networks, storage, servers, and other computing resources are provided to client companies. Clients can run their application and have administrative rights to use these resources but do not manage underlying infrastructure. Clients have to pay for usage of infrastructure. A good example of this is **Amazon.com**'s Web services. **Amazon.com** has developed impressive technology infrastructure that includes data centers. Other companies can use **Amazon.com**'s cloud services on a pay-per-use-basis without having to make similar investments. Similar services are offered by all major cloud providers such as IBM, Microsoft, Google, and so on.

We should note that there is considerable confusion and overlap in the use of cloud terminology. For example, some vendors also add information as a service (IaaS), which is an extension of DaaS. Clearly, *this* IaaS is different from infrastructure as a service described earlier. Our goal here is to just recognize that there are varying

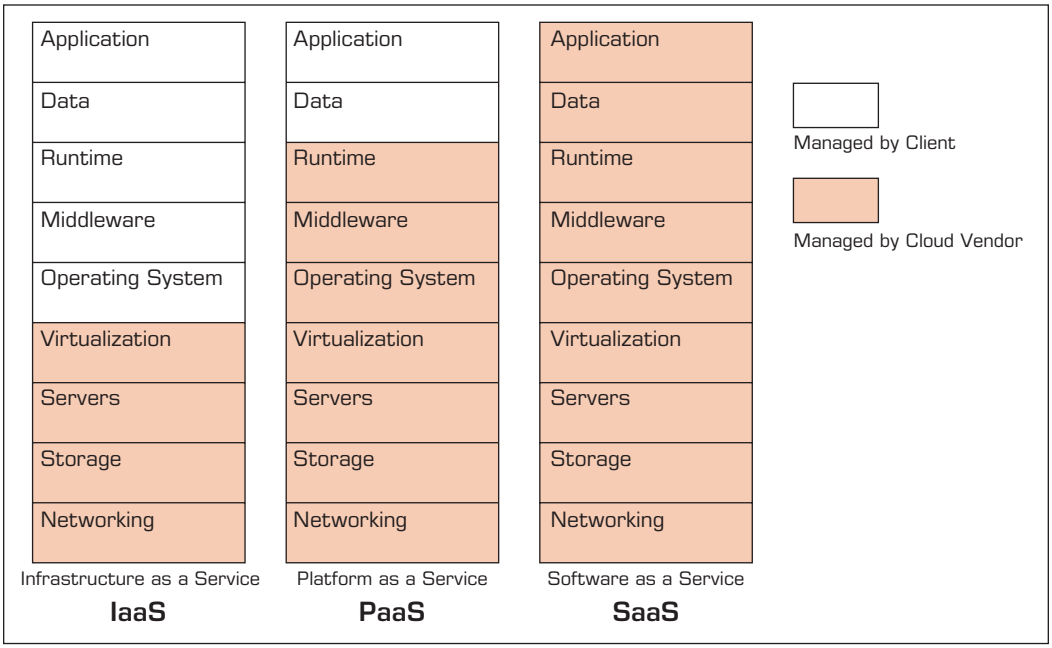


FIGURE 9.13 **Technology Stack as a Service for Different Types of Cloud Offerings.**

degrees of services that an organization can subscribe to in order to manage the analytics applications. Figure 9.13 highlights the level of service subscriptions a client uses in each of the three major types of cloud offerings. SaaS is clearly the highest level of cloud service that a client may get. For example, in using Office 365, an organization is using the software as a service. The client is only responsible for bringing in the data. Many of the analytics as a service application fall in this category as well. Further, several analytics as a service provider may in turn use clouds such as Amazon's AWS or Microsoft Azure to provide their services to the end users. We will see examples of such services shortly.

Essential Technologies for Cloud Computing

VIRTUALIZATION Virtualization is the creation of a virtual version of something like an operating system or server. A simple example of virtualization is the logical division of a hard drive to create two separate hard drives in a computer. Virtualization can be in all three areas of computing:

- **Network virtualization:** It is the splitting of available bandwidth into channels, which disguises complexity of the network by dividing it into manageable parts. Then each bandwidth can be allocated to a particular server or device in real time.
- **Storage virtualization:** It is the pooling of physical storage from multiple network storage devices into a single storage device that can be managed from a central console.
- **Server virtualization:** It is the masking of physical servers from server users. Users don't have to manage the actual servers or understand complicated details of server resources.

This difference in the level of virtualization directly relates to which cloud service one employs.

Application Case 9.9 illustrates an application of cloud technologies that enable a mobile application and allow for significant reduction in information miscommunication.

Application Case 9.9

Major West Coast Utility Uses Cloud-Mobile Technology to Provide Real-Time Incident Reporting

Historical communication between utilities and first responders has been by phone calls or two-way radios. Some of these are with first responders on the scene, and some with dispatch or other units of the first responder organization. When a member of the public sees an incident on the field, they usually just call 911, which is routed to first responders. Dispatch centers route the closest first responder to the field, who then call back to the center either on their radios or cell phones to let them know the actual status. The dispatch centers then call the incident in to the appropriate utility, who then sends their own team to the field for further resolution. This also requires that the exact location be conveyed to the dispatch center from the field, and from the former to utility—particularly challenging if the incident location is not at a specific address (e.g., along a freeway, across open land, etc.). The utility also needs to let the dispatch center know the status of their own crew. This information must also be relayed to the first responders on the field. Much of this process relies on information being communicated orally and then forwarded to one or more recipients, with information also flowing back and forth along the same chain. All of this can result in garbled communication and/or incomplete messages, which can eat away precious minutes or even hours in emergencies.

A major West Coast Utility, a leader in using technology to address traditional problems, determined that many of these challenges can be addressed through better information sharing in a timelier manner using cloud-mobile technology. Their territory encompassed densely populated cities to far-flung rural communities with intervening miles of desert, national parks, and more.

Recognizing that most first responders have a smartphone or tablet, the utility selected Connixt's iMarq™ mobile suite to provide a simple-to-use mobile app that allows first responders to advise the utility of any incident in the field. The technology also keeps the first responders apprised of the utility's response status with respect to the incident.

With a targeted base of over 20,000 first responders spread across the entire territory, lowering barriers to adoption was an especially important factor. "Improving communication with groups that are outside your organization is historically difficult," says G. Satish, cofounder and CEO, Connixt. "For this deployment, the focus on simplicity is the key to its success."

First responders are invited to download and self-register the app, and once the utility grants access rights, they can report incidents using their own tablets or smartphones. The first responder simply uses a drop-down menu to pick from a list of preconfigured incidents, taps an option to indicate if they will wait at the scene, and attach photographs with annotations—all with a few touches on their device. The utility receives notification of the incident, reviews the time and geostamped information (no more mixed-up addresses), and updates their response. This response (which may be a truck roll) is sent to the first responders and maintained in the app.

The simplicity of the solution makes it easy for the first responders. They use their own phone or tablet, communicate in a way they are used to, and provide needed information simply and effectively. They can see the utility updates (such as the current status of the truck that was sent). Missed or garbled phone messages are minimized. Options such as recording voice memos, using speech-to-text and more, are also available.

Cloud technology has been particularly useful in this case—deployment is faster without issues related to hardware procurement, installation, and appropriate backups. Connixt's cloud-based Mobile Extension Framework (MXF™) is architected for rapid configuration and deployment—configuration is completed in the cloud, and, once configured, the apps are ready for download and deployment. More importantly, MXF enables easy modifications to forms and processes—for example, if the utility needs to add additional options to the incident drop-down, they simply add this once in MXF. Within minutes the option is now available on the field for all users. Figure 9.14 illustrates this architecture.

There are further benefits from a system that leverages ubiquitous cloud and mobile technologies. Because all of the business logic and configurations are stored in the cloud, the solution itself can act as a stand-alone system for customers who have no back-end systems—very important in the context of small and medium businesses (SMBs). And for those with back-end systems, the connectivity is seamless through Web services and the back-end system serves as the

(*Continued*)

Application Case 9.9 (Continued)

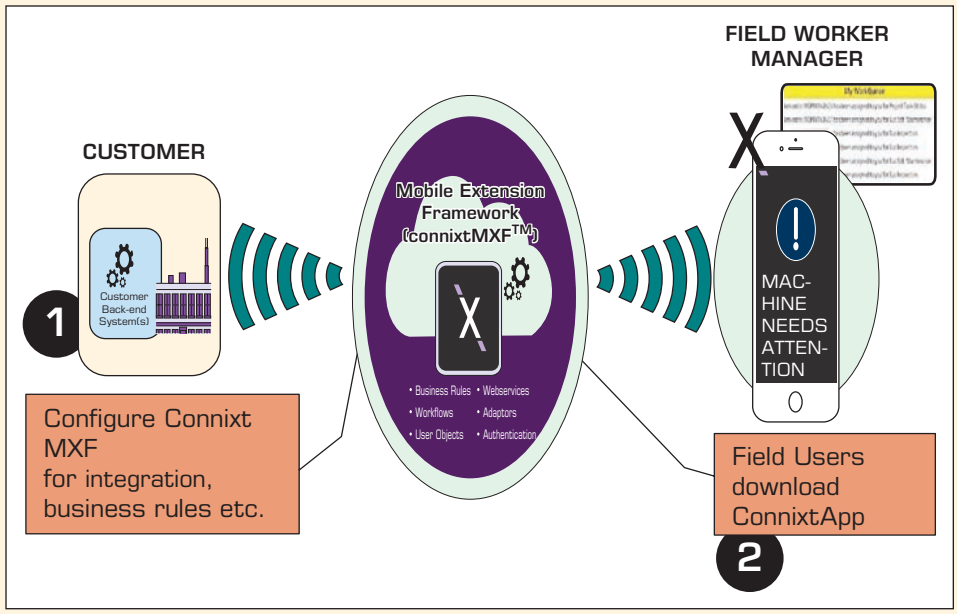


FIGURE 9.14 Interconnections between workers and technology in a cloud analytics application.

system of record. This additionally helps businesses adopt technology in a phased manner—starting with a noninvasive, standalone system with minimal internal IT impact while automating field operations, and then moving toward back-end system integration.

On the other hand, the mobile apps are themselves system agnostic—they communicate using standard Web services and the end device can be Android or iOS and smartphone or tablet. Thus, irrespective of the device used, all communication, business logic, and algorithms are standardized across platforms/devices. As native apps across all devices, iMarq leverages standard technology that is provided by the device manufacturers and the OS vendors. For example, using native maps applications allows the apps to benefit from improvements made by the platform vendors; thus, as maps become more accurate, the end users of the mobile apps also benefit from these advances.

Finally, for successful deployments, enterprise cloud-mobile technology has to be heavily user-centric. The look and feel must be geared to user-comfort, much as users expect from any mobile app they use. Treating the business user as an app consumer meets their standard expectations of an intuitive app that immediately saves them time and effort. This approach is essential to ensuring successful adoption.

The utility now has better information from first responders, as information is directly shared from the field (not through a dispatcher or other third party), pictures are available, and there is geo- and time stamping. Garbled phone messages are avoided. The two-way communication between utility and the first responder in the field is improved. Historical records of the incidents are kept.

The utility and the first responders are now more unified in their quick and complete responses to incidents, improving service to the public. By tightening ties with first responders (police and fire department personnel), the public is served with a better coordinated and superior response for incidents that are discovered by first responders.

Questions for Discussion

1. How does cloud technology impact enterprise software for small and mid-size businesses?
2. What are some of the areas where businesses can use mobile technology?
3. What types of businesses are likely to be the forerunners in adopting cloud-mobile technology?
4. What are the advantages of cloud-based enterprise software instead of the traditional on-premise model?
5. What are the likely risks of cloud versus traditional on-premise applications?

Source: Used with permission from G Satish, Connixit, Inc.

Cloud Deployment Models

Cloud services can be acquired in several ways, from building an entirely private infrastructure to sharing with others. The following three models are the most common.

Private cloud: This can also be called internal cloud or corporate cloud. It is a more secure form of cloud service than public clouds like Microsoft Azure and Amazon Web Services. It is operated solely for a single organization having a mission critical workload and security concerns. It provides the same benefits as a public cloud-like service, scalability, changing computing resources on demand, and so on. Companies that have a private cloud have direct control over their data and applications. The disadvantage of having a private cloud is the cost of maintaining and managing the cloud because on-premise IT staff are responsible for managing it.

Public cloud: In this model the subscriber uses the resources offered by service providers over the Internet. The cloud infrastructure is managed by the service provider. The main advantage of this public cloud model is saving time and money in setting up hardware and software required to run their business. Examples of public clouds are Microsoft Azure, Google Cloud Platform, and Amazon AWS.

Hybrid cloud: The hybrid cloud gives businesses great flexibility by moving workloads between private and public clouds. For example, a company can use hybrid cloud storage to store its sales and marketing data, and then use a public cloud platform like Amazon Redshift to run analytical queries to analyze its data. The main requirement is network connectivity and API (application program interface) compatibility between the private and public cloud.

Major Cloud Platform Providers in Analytics

This section first identifies some key cloud players that provide the infrastructure for analytics as a service, as well as selected analytics functionalities. Then we also mention representative analytics-as-a-service offerings that may even run on these cloud platforms.

Amazon Elastic Beanstalk: Amazon Elastic Beanstalk is a service offered from Amazon Web Services. It can deploy, manage, and scale Web applications. It supports the following programming languages: Java, Ruby, Python, PHP, and .NET on servers like Apache HTTP, Apache Tomcat, and IIS. A user has to upload the code for the application, and Elastic Beanstalk handles the deployment of the application, load balancing, and autoscaling and monitors the health of the application. So the user can focus on building Web sites, mobile applications, API backend, content management systems, SaaS, and so on, while the applications and infrastructure to manage them are taken care of by Elastic Beanstalk. A user can use Amazon Web Services or an integrated development environment like Eclipse or Visual Studio to upload their application. A user has to pay for AWS resources needed to store and run the applications.

IBM Cloud: IBM Cloud is a cloud platform that allows a user to build apps using many open source computer technologies. Users can also deploy and manage hybrid applications using the software. With IBM Watson, whose services are available on IBM Cloud, users can now create next-generation cognitive applications that can discover, innovate, and make decisions. IBM Watson services can be used for analyzing emotions and synthesizing natural-sounding speech from text. Watson uses the concept of cognitive computing to analyze text, video, and images. It supports programming languages like Java, Go, PHP, Ruby, and Python.

Microsoft Azure: Azure is a cloud platform created by Microsoft to build, deploy, and manage applications and services through a network of Microsoft data centers.

It serves as both PaaS and IaaS and offers many solutions such as analytics, data warehousing, remote monitoring, and predictive maintenance.

Google App Engine: Google App Engine is Google's Cloud computing platform used for developing and hosting applications. Managed by Google's data centers, it supports developing apps in Python, Java, Ruby, and PHP programming languages. The big query environment offers data warehouse services through the cloud.

Openshift: Openshift is Red Hat's cloud application platform based on a PaaS model. Through this model, application developers can deploy their applications on the cloud. There are two different models available for openshift. One serves as a public PaaS and the other serves as a private PaaS. Openshift Online is Red Hat's public PaaS that offers development, build, hosting, and deployment of applications in the cloud. The private PaaS, openshift Enterprise, allows development, build, and deployment of applications on an internal server or a private cloud platform.

Analytics as a Service (AaaS)

Analytics and data-based managerial solutions—the applications that query data for use in business planning, problem solving, and decision support—are evolving rapidly and being used by almost every organization. Enterprises are being flooded with information, and getting insights from this data is a big challenge for them. Along with that, there are challenges related to data security, data quality, and compliance. AaaS is an extensible analytical platform using a cloud-based delivery model where various BI and data analytics tools can help companies in better decision making and get insights from their huge amount of data. The platform covers all functionality aspects from collecting data from physical devices to data visualization. AaaS provides an agile model for reporting and analytics to businesses so they can focus on what they do best. Customers can either run their own analytical applications in the cloud or they can put their data on the cloud and receive useful insights.

AaaS combines aspects of cloud computing with Big Data analytics and empowers data scientists and analysts by allowing them to access centrally managed information data sets. They can now explore information data sets more interactively and discover richer insights more rapidly, thus erasing many of the delays that they may face while discovering data trends. For example, a provider might offer access to a remote analytics platform for a fee. This allows the client to use analytics software for as long as it is needed. AaaS is a part of SaaS, PaaS, and IaaS, thus helping IT significantly reduce costs and compliance risk, while increasing productivity of users.

AaaS in the cloud has economies of scale and scope by providing many virtual analytical applications with better scalability and higher cost savings. With growing data volumes and dozens of virtual analytical applications, chances are that more of them leverage processing at different times, usage patterns, and frequencies.

Data and text mining is another very promising application of AaaS. The capabilities that a service orientation (along with cloud computing, pooled resources, and parallel processing) brings to the analytics world can also be used for large-scale optimization, highly complex multicriteria decision problems, and distributed simulation models. Next we identify selected cloud-based analytics offerings.

Representative Analytics as a Service Offerings

IBM CLOUD IBM is making all of its analytics offerings available through its cloud. IBM Cloud offers several categories of analytics and AI. For example, IBM Watson Analytics integrates most of the analytics features and capabilities that can be built and deployed through their cloud. In addition, IBM Watson Cognitive has been a major cloud-based

offering that employs text mining and deep learning at a very high level. It was introduced earlier in the context of text mining.

MINEMYTEXT.COM One of the areas of major growth in analytics is text mining. Text mining identifies high-level topics of documents, infers sentiments from reviews, and visualizes the document or term/concept relationships, as covered in the text mining chapter. A start-up called **MineMyText.com** offers these capabilities in the cloud through their Web site.

SAS VIYA SAS Institute is making its analytics software offering available on demand through the cloud. Currently, SAS Visual Statistics is only available as a cloud service and is a competitor of Tableau.

TABLEAU Tableau, a major visualization software that was introduced in the context of descriptive analytics, is also available through the cloud.

SNOWFLAKE Snowflake is a cloud-based data warehouse solution. Users can bring together their data from multiple sources as one source and analyze it using Snowflake.

Illustrative Analytics Applications Employing the Cloud Infrastructure

In this section we highlight several cloud analytics applications. We present them as one section as opposed to individual Application Cases.

Using Azure IOT, Stream Analytics, and Machine Learning to Improve Mobile Health Care Services

People are increasingly using mobile applications to keep track of the amount of exercise they do every day and maintain their health history as well. Zion China, which is a provider of mobile healthcare services, has come up with an innovative health monitoring tool that gathers data about health problems such as glucose levels, blood pressure, diet, medication, and exercise of their users and help them improve their quality of life by giving them suggestions on how to manage their health and prevent or cure illness on a daily basis.

The huge volume of real-time data presented scalability and data management problems, so the company collaborated with Microsoft to take advantage of Stream Analytics, Machine Learning, IOT solution and Power BI, which also improved data security and analysis. Zion China was completely dependent on traditional BI with data being collected from various devices or cloud. Using a cloud-based analytics architecture, Zion was able to add several features, speed, and security. They added an IoT hub to the front end for better transmission of data from device to cloud. The data is first transferred from the device to a mobile application via Bluetooth and then to an IoT hub via HTTPS and AMQP. Stream Analytics helps in processing the real time gathered in the IoT hub, and generates insights and useful information, which is further streamed to an SQL database. They use Azure Machine Learning to generate predictive models on diabetes patient data and improve the analysis and prediction levels. Power BI provides simple and easy visualization of data insights achieved from analysis to the users.

Sources: "Zion China Uses Azure IoT, Stream Analytics, and Machine Learning to Evolve Its Intelligent Diabetes Management Solution" at **www.codeproject.com/Articles/1194824/Zion-China-uses-Azure-IoT-Stream-Analytics-and-M** (accessed October 2018) and **https://microsoft.github.io/techcasestudies/iot/2016/12/02/IoT-ZionChina.html** (accessed October 2018).

Gulf Air Uses Big Data to Get Deeper Customer Insight

Gulf Air is the national carrier of Bahrain. It is a major international carrier with 3,000 employees, serving 45 cities in 24 countries across three continents. Gulf Air is an industry leader in providing traditional Arabian hospitality to customers. To learn more about how their customers felt about their hospitality services, the airline wanted to know what their customers were saying on social media about the airline's hospitality. The challenge was analyzing all the comments and posts from their customers, as there were hundreds of thousands of posts every day. Monitoring these posts manually would be a time-consuming and daunting task and would also be prone to human error.

Gulf Air wanted to automate this task and analyze the data to learn of the emerging market trends. Along with that, the company wanted a robust infrastructure to host such a social media monitoring solution that would be available around the clock and agile across geographical boundaries.

Gulf Air developed a sentiment analysis solution, "Arabic Sentiment Analysis," that analyzes English and Arabic social media posts. The Arabic Sentiment Analysis tool is based on Cloudera's distribution of Hadoop Big Data framework. It runs on Gulf Air's private cloud environment and also uses the Red Hat JBoss Enterprise Application platform. The private cloud holds about 50 terabytes of data, and the Arabic Sentiment Analysis tool can analyze thousands of posts on social media, providing sentiment results in minutes.

Gulf Air achieved substantial cost savings by putting the "Arabic Sentiment Analysis" application on the company's existing private cloud environment as they didn't need to invest in setting up the infrastructure for deploying the application. "Arabic Sentiment Analysis" helps Gulf Air in deciding promotions and offers for their passengers on a timely basis and helps them stay ahead of their competitors. In case the master server fails, the airline created "ghost images" of the server that can be deployed quickly, and the image can start functioning in its place. The Big Data solution quickly and efficiently captures posts periodically and transforms them into reports, giving Gulf Air up-to-date views of any change in sentiment or shifts in demand, enabling them to respond quickly. Insights from the Big Data solution have had a positive impact on the work performed by the employees of Gulf Air.

Sources: **RedHat.com**. (2016). "Gulf Air Builds Private Cloud for Big Data Innovation with Red Hat Technologies." **www.redhat.com/en/about/press-releases/gulf-air-builds-private-cloud-big-data-innovation-red-hat-technologies** (accessed October 2018); **RedHat.com**. (2016). "Gulf Air's Big Data Innovation Delivers Deeper Customer Insight." **www.redhat.com/en/success-stories** (accessed October 2018); **ComputerWeekly.com**. (2016). "Big-Data and Open Source Cloud Technology Help Gulf Air Pin Down Customer Sentiment." **www.computerweekly.com/news/450297404/Big-data-and-open-source-cloud-technology-help-Gulf-Air-pin-down-customer-sentiment** (accessed October 2018).

Chime Enhances Customer Experience Using Snowflake

Chime, a banking option, offers a Visa debit card, FDIC-insured spending and savings account, and a mobile application app that makes banking easier for people. Chime wanted to learn about their customer engagement. They wanted to analyze data across their mobile, Web, and backend platforms to help enhance the user experience. However, pulling and aggregating data from multiple sources such as ad services from Facebook and Google and events from other third-party analytics tools like JSON (JavaScript Object Notation) docs, was a laborious task. They wanted a solution that could aggregate data from these multiple sources and analyze the data set. Chime needed a solution that could process JSON data sources and query them using standard SQL database tables.

Chime started using Snowflake Elastic Data Warehouse solution. Snowflake pulled data from all 14 data sources of chime, including data like JSON docs from applications.

Snowflake helped Chime analyze JSON data quickly to enhance member services and provide a more personalized banking experience to customers.

Source: Based on **Snowflake.net.** (n.d.). Chime delivers personalized customer experience using Chime. **http://www.snowflake.net/product** (accessed Oct 2018).

We are entering the "petabyte age," and traditional data and analytics approaches are beginning to show their limits. Cloud analytics is an emerging alternative solution for large-scale data analysis. Data-oriented cloud systems include storage and computing in a distributed and virtualized environment. A major advantage of these offerings is the rapid diffusion of advanced analysis tools among the users, without significant investment in technology acquisition. These solutions also come with many challenges, such as security, service level, and data governance. A number of concerns have been raised about cloud computing, including loss of control and privacy, legal liabilities, cross-border political issues, and so on. According to Cloud Security Alliance, the top three security threats in the cloud are data loss and leakage, hardware failure of equipment, and an insecure interface. All the data in the cloud is accessible by the service provider, so the service provider can unknowingly or deliberately alter the data or can pass the data to a third party for purposes of law without asking the company. Research is still limited in this area. As a result, there is ample opportunity to bring analytical, computational, and conceptual modeling into the context of service science, service orientation, and cloud intelligence. Nonetheless, cloud computing is an important initiative for an analytics professional to watch as it is a fast-growing area.

SECTION 9.9 REVIEW QUESTIONS

1. Define *cloud computing*. How does it relate to PaaS, SaaS, and IaaS?
2. Give examples of companies offering cloud services.
3. How does cloud computing affect BI?
4. How does DaaS change the way data is handled?
5. What are the different types of cloud platforms?
6. Why is AaaS cost-effective?
7. Name at least three major cloud service providers.
8. Give at least three examples of analytics-as-a-service providers.

9.10 LOCATION-BASED ANALYTICS FOR ORGANIZATIONS

Thus far, we have seen many examples of organizations employing analytical techniques to gain insights into their existing processes through informative reporting, predictive analytics, forecasting, and optimization techniques. In this section, we learn about a critical emerging trend—incorporation of location data in analytics. Figure 9.15 gives our classification of location-based analytic applications. We first review applications that make use of static location data that is usually called *geospatial data*. We then examine the explosive growth of applications that take advantage of all the location data being generated by today's devices. This section first focuses on analytics applications that are being developed by organizations to make better decisions in managing operations, targeting customers, promotions, and so forth. Then we will also explore analytics applications that are being developed to be used directly by a consumer, some of which also take advantage of the location data.

Geospatial Analytics

A consolidated view of the overall performance of an organization is usually represented through the visualization tools that provide actionable information. The information may include current and forecasted values of various business factors and

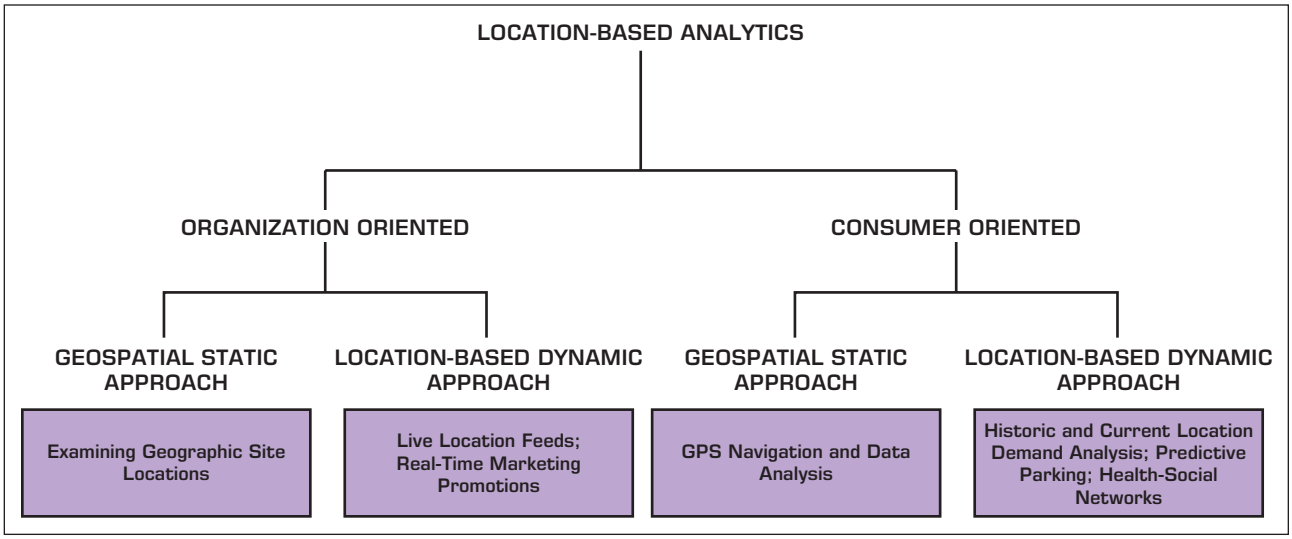


FIGURE 9.15 **Classification of Location-Based Analytics Applications.**

key performance indicators (KPIs). Looking at the KPIs as overall numbers via various graphs and charts can be overwhelming. There is a high risk of missing potential growth opportunities or not identifying the problematic areas. As an alternative to simply viewing reports, organizations employ visual maps that are geographically mapped and based on the traditional location data, usually grouped by postal codes. These map-based visualizations have been used by organizations to view the aggregated data and get more meaningful location-based insights. The traditional location-based analytic techniques using geocoding of organizational locations and consumers hamper the organizations in understanding "true location-based" impacts. Locations based on postal codes offer an aggregate view of a large geographic area. This poor granularity may not help pinpoint the growth opportunities within a region, as the location of target customers can change rapidly. Thus, an organization's promotional campaigns may not target the right customers if it is based on postal codes. To address these concerns, organizations are embracing location and spatial extensions to analytics. The addition of location components based on latitudinal and longitudinal attributes to the traditional analytical techniques enables organizations to add a new dimension of "where" to their traditional business analyses, which currently answers the questions of "who," "what," "when," and "how much."

Location-based data are now readily available from **geographic information systems (GIS)**. These are used to capture, store, analyze, and manage data linked to a location using integrated sensor technologies, global positioning systems installed in smartphones, or through RFID deployments in the retail and healthcare industries.

By integrating information about the location with other critical business data, organizations are now creating location intelligence. Location intelligence is enabling organizations to gain critical insights and make better decisions by optimizing important processes and applications. Organizations now create interactive maps that further drill down to details about any location, offering analysts the ability to investigate new trends and correlate location-specific factors across multiple KPIs. Analysts can now pinpoint trends and patterns in revenue, sales, and profitability across geographical areas.

By incorporating demographic details into locations, retailers can determine how sales vary by population level and proximity to other competitors; they can assess the demand and efficiency of supply-chain operations. Consumer product companies can identify the specific needs of customers and customer complaint locations and easily trace them back to the products. Sales reps can better target their prospects by analyzing their geography.

A company that is the market leader in providing GIS data is ESRI (**esri.com**). ESRI licenses its ArcGIS software to thousands of customers including commercial, government, and the military. It would take a book or more to highlight applications of ESRI's GIS database and software! Another company **grindgis.com** identifies over 60 categories of GIS applications (**http://grindgis.com/blog/gis-applications-uses** (accessed October 2018)). A few examples that have not been mentioned yet include the following:

Agricultural applications: By combining location, weather, soil, and crop-related data, very precise irrigation and fertilizer applications can be planned. Examples include companies such as **sstsoftware.com** and **sensefly.com** (they combine GIS and the latest information collected through drones, another emerging technology).

Crime analysis: Superimposition of crime data including date, time, and type of crime onto the GIS data can provide significant insights into crime patterns and police staffing.

Disease spread prediction: One of the first known examples of descriptive analytics is the analysis of the cholera outbreak in London in 1854. Dr. John Snow plotted the cases of cholera on a map and was able to refute the theory that the cholera outbreak was being caused by bad air. The map helped him pinpoint the outbreak to a bad water well (**TheGuardian.com**, 2013). We have come a long way from needing to plot maps manually, but the idea of being able to track and then predict outbreaks of diseases, such as the flu, using GIS and other data has become a major field in itself. Application Case 9.7 gave an example of using social media data along with GIS data to pinpoint flu trends.

In addition, with location intelligence, organizations can quickly overlay weather and environmental effects and forecast the level of impact on critical business operations. With technology advancements, geospatial data is now being directly incorporated in enterprise data warehouses. Location-based in-database analytics enable organizations to perform complex calculations with increased efficiency and get a single view of all the spatially oriented data, revealing hidden trends and new opportunities. For example, Teradata's data warehouse supports the geospatial data feature based on the SQL/MM standard. The geospatial feature is captured as a new geometric data type called ST_GEOMETRY. It supports a large spectrum of shapes, from simple points, lines, and curves to complex polygons in representing the geographic areas. They are converting the nonspatial data of their operating business locations by incorporating the latitude and longitude coordinates. This process of geocoding is readily supported by service companies like NAVTEQ and Tele Atlas, which maintain worldwide databases of addresses with geospatial features and make use of address-cleansing tools like Informatica and Trillium, which support mapping of spatial coordinates to the addresses as part of extract, transform, and load functions.

Organizations across a variety of business sectors are employing geospatial analytics. We will review some examples next. Application Case 9.10 provides an example of how location-based information was used in making site selection decisions in expanding a company's footprint. Application Case 9.11 illustrates another application that goes beyond just the location decision.

Application Case 9.10 Great Clips Employs Spatial Analytics to Shave Time in Location Decisions

Great Clips, one of the world's largest and fastest-growing hair salons, has more than 3,000 salons throughout the United States and Canada. Great Clips franchise success depends on a growth strategy that is driven by rapidly opening new stores in the right locations and markets. The company needed to analyze the locations based on the requirements for a potential customer base, demographic trends, and sales impact on existing franchises in the target location. Choosing a good site is of utmost importance. The current processes took a long time to analyze a single site and a great deal of labor requiring intensive analyst resources to manually assess the data from multiple data sources.

With thousands of locations to analyze each year, the delay was risking the loss of prime sites to competitors and was proving expensive; Great Clips employed external contractors to cope with the delay. The company created a site-selection workflow application to evaluate the new salon site locations by using the geospatial analytical capabilities of Alteryx. A new site location was evaluated by its drive-time proximity and convenience for serving all the existing customers of the Great Clips network in the area. The Alteryx-based solution also enabled evaluation of each new location based on demographics and consumer behavior data, aligning with existing Great Clips customer profiles and the potential impact of new site revenue on the existing sites. As a result of using location-based analytic techniques, Great Clips was able to reduce the time to assess new locations by nearly 95%. The labor-intensive analysis was automated and developed into a data collection analysis, mapping, and reporting application that could be easily used by the nontechnical real estate managers. Furthermore, it enabled the company to implement proactive predictive analytics for a new franchise location, as the whole process now took just a few minutes.

Questions for Discussion

1. How is geospatial analytics employed at Great Clips?
2. What criteria should a company consider in evaluating sites for future locations?
3. Can you think of other applications where such geospatial data might be useful?

Source: Based on **Alteryx.com**. Great Clips. **alteryx.com/sites/default/files/resources/files/case-study-great-chips.pdf** (accessed Sept 2018).

Application Case 9.11 Starbucks Exploits GIS and Analytics to Grow Worldwide

One of the key challenges for any organization that is trying to grow its presence is deciding the location of its next store. Starbucks faces the same question. To identify new store locations, more than 700 Starbucks employees (referred to as partners) in 15 countries use an ArcGIS-based market planning and BI solution called Atlas. Atlas provides partners with workflows, analysis, and store performance information so that local partners in the field can make decisions when identifying new business opportunities.

As reported in multiple sources, Atlas is employed by local decision makers to understand the population trends and demand. For example, in China, there are over 1,200 Starbucks stores, and the company is opening a new store almost every day. Information such as trade areas, retail clusters and generators, traffic, and demographics is important in deciding the next store's location. After analyzing a new market and neighborhood, a manager can look at specific locations by zooming into an area in the city and identifying where three new office towers may be completed over the next 2 months, for example. After viewing this area on the map, a workflow window can be created that will help the manager move the new site through approval, permitting, construction, and eventually opening.

By integrating weather and other local data, one can also better manage demand and supply-chain operations. Starbucks is integrating its enterprise business systems with its GIS solutions in Web services to see the world and its business in new ways. For example, Starbucks integrates AccuWeather's forecasted real-feel temperature data. This forecasted temperature data can help localize marketing efforts. If a really hot week in Memphis is forthcoming, Starbucks analysts can select a group of coffee houses and get detailed information on past and future weather patterns, as well as store characteristics. This knowledge can be used to design a localized promotion for Frappuccinos, for example, helping Starbucks anticipate what its customers will be wanting a week in advance.

Major events also have an impact on coffee houses. When 150,000 people descended on San Diego for the Pride Parade, local baristas served a lot of customers. To ensure the best possible customer experience, Starbucks used this local event knowledge to plan staffing and inventory at locations near the parade.

Questions for Discussion

1. What type of demographics and GIS information would be relevant for deciding on a store location?
2. It has been mentioned that Starbucks encourages its customers to use its mobile app. What type of information might the company gather from the app to help it better plan operations?
3. Will the availability of free Wi-Fi at Starbucks stores provide any information to Starbucks for better analytics?

Sources: Digit.HBS.org. (2015). "Starbucks: Brewing up a Data Storm!" **https://digit.hbs.org/submission/starbucks-brewing-up-a-data-storm/** (accessed October 2018); Wheeler, C. (2014). "Going Big with GIS." **www.esri.com/esri-news/arcwatch/0814/going-big-with-gis** (accessed October 2018); **Blogs.ESRI.com**. "From Customers to CxOs, Starbucks Delivers World-Class Service." (2014). **https://blogs.esri.com/esri/ucinsider/2014/07/29/starbucks/** (accessed October 2018).

In addition to the retail transaction analysis applications highlighted here, there are many other applications of combining geographic information with other data being generated by an organization. For example, network operations and communication companies often generate massive amounts of data every day. The ability to analyze the data quickly with a high level of location-specific granularity can better identify the customer churn and help in formulating strategies specific to locations for increasing operational efficiency, quality of service, and revenue.

Geospatial analysis can enable communication companies to capture daily transactions from a network to identify the geographic areas experiencing a large number of failed connection attempts of voice, data, text, or Internet. Analytics can help determine the exact causes based on location and drill down to an individual customer to provide better customer service. You can see this in action by completing the following multimedia exercise.

A Multimedia Exercise in Analytics Employing Geospatial Analytics

Teradata University Network includes a BSI video on the case of dropped mobile calls. Please watch the video that appears on YouTube at the following link: **www.teradatauniversitynetwork.com/Library/Samples/BSI-The-Case-of-the-Dropped-Mobile-Calls** (accessed October 2018).

A telecommunication company launches a new line of smartphones and faces problems with dropped calls. The new rollout is in trouble, and the northeast region is the worst hit region as they compare effects of dropped calls on the profits for the geographic region. The company hires BSI to analyze the problems arising due to defects in smartphone handsets, tower coverage, and software glitches. The entire northeast region data

is divided into geographic clusters, and the company solves the problem by identifying the individual customer data. The BSI team employs geospatial analytics to identify the locations where network coverage was leading to dropped calls and suggests installing a few additional towers where unhappy customers are located.

After the video is complete, you can see how the analysis was prepared at: **slideshare.net/teradata/bsi-teradata-the-case-of-the-dropped-mobile-calls** (accessed October 2018).

This multimedia excursion provides an example of a combination of geospatial analytics along with Big Data analytics that assist in better decision making.

Real-Time Location Intelligence

Many devices in use by consumers and professionals are constantly sending out their location information. Cars, buses, taxis, mobile phones, cameras, and personal navigation devices all transmit their locations thanks to network-connected positioning technologies such as GPS, Wi-Fi, and cell tower triangulation. Millions of consumers and businesses use location-enabled devices for finding nearby services, locating friends and family, navigating, tracking assets and pets, dispatching, and engaging in sports, games, and hobbies. This surge in location-enabled services has resulted in a massive database of historical and real-time streaming location information. It is, of course, scattered and not very useful by itself. The automated data collection enabled through capture of cell phones and Wi-Fi hotspot access points presents an interesting new dimension in nonintrusive market research, data collection, and, of course, microanalysis of such massive data sets.

By analyzing and learning from these large-scale patterns of movement, it is possible to identify distinct classes of behaviors in specific contexts. This approach allows a business to better understand its customer patterns and make more informed decisions about promotions, pricing, and so on. By applying algorithms that reduce the dimensionality of location data, one can characterize places according to the activity and movement between them. From massive amounts of high-dimensional location data, these algorithms uncover trends, meaning, and relationships to eventually produce human-understandable representations. It then becomes possible to use such data to automatically make intelligent predictions and find important matches and similarities between places and people.

Location-based analytics finds its application in consumer-oriented marketing applications. Many companies are now offering platforms to analyze location trails of mobile users based on geospatial data obtained from the GPS and target tech-savvy customers with coupons on their smartphones as they pass by a retailer. This illustrates the emerging trend in the retail space where companies are looking to improve efficiency of marketing campaigns—not just by targeting every customer based on real-time location, but by employing more sophisticated predictive analytics in real time on consumer behavioral profiles to find the right set of consumers for advertising campaigns.

Yet another extension of location-based analytics is to use augmented reality. In 2016, Pokémon GO became a market sensation. It is a location-sensing augmented reality-based game that encourages users to claim virtual items from select geographic locations. The user can start anywhere in a city and follow markers on the app to reach a specific item. Virtual items are visible through the app when the user points a phone's camera toward the virtual item. The user can then claim this item. Business applications of such technologies are also emerging. For example, an app called Candybar allows businesses to place these virtual items on a map using Google Maps. The placement of this item can be fine-tuned using Google's Street View. Once all virtual items have been configured with the information and location, the business can submit items, which are then visible to the user in real time. Candybar also provides usage analytics to the business to enable

better targeting of virtual items. The virtual reality aspect of this app improves the experience of users, providing them with a "gaming" environment in real life. At the same time, it provides a powerful marketing platform for businesses to reach their customers.

As is evident from this section, location-based analytics and ensuing applications are perhaps the most important front in the near future for organizations. A common theme in this section was the use of operational or marketing data by organizations. We will next explore analytics applications that are directly targeted at users and sometimes take advantage of location information.

Analytics Applications for Consumers

The explosive growth of the apps industry for smartphone platforms (iOS, Android, Windows, and so forth) and the use of analytics are creating tremendous opportunities for developing apps where the consumers use analytics without ever realizing it. These apps differ from the previous category in that these are meant for direct use by a consumer, as opposed to an organization that is trying to mine a consumer's usage/purchase data to create a profile for marketing specific products or services. Predictably, these apps are meant for enabling consumers to make better decisions by employing specific analytics. We highlight two of these in the following examples.

- Waze, a social Web app that assists users in identifying a navigation path and alerts users about potential issues such as accidents, police checkpoints, speed traps, and construction, based on other users' inputs, has become a very popular navigation app. Google acquired this app a few years ago and has enhanced it further. This app is an example of aggregating user-generated information and making it available for customers.
- Many apps allow users to submit reviews and ratings for businesses, products, and so on, and then present those to the users in an aggregated form to help them make choices. These apps can also be identified as apps based on social data that are targeted at consumers where the data are generated by the consumers. One of the more popular apps in this category is Yelp. Similar apps are available all over the world.
- Another transportation-related app that uses predictive analytics, ParkPGH, has been deployed since about 2010 in Pittsburgh, Pennsylvania. Developed in collaboration with Carnegie Mellon University, this app includes predictive capabilities to estimate parking availability. ParkPGH directs drivers to parking lots in areas where parking is available. It calculates the number of parking spaces available in several garages in the cultural arts district of Pittsburgh. Available spaces are updated every 30 seconds, keeping the driver as close to the current availability as possible. Depending on historical demand and current events, the app is able to predict parking availability and provide information on which lots will have free space by the time the driver reaches the destination. The app's underlying algorithm uses data on current events around the area—for example, a basketball game—to predict an increase in demand for parking spaces later that day, thus saving the commuters valuable time searching for parking spaces in the busy city. Success of this app has led to a proliferation of parking apps that work in many major cities and allow a user to book a parking space in advance, recharge the meter, even bid for a parking space, etc. Both iPhone app store and Google Play store include many such apps.

Analytics-based applications are emerging not just for fun and health, but also to enhance one's productivity. For example, Google's e-mail app called Gmail analyzes billions of e-mail transactions and develops automated responses for e-mails. When a

user receives an e-mail and reads it in her Gmail app, the app also recommends short responses for the e-mail at hand that a user can select and send to the original sender.

As is evident from these examples of consumer-centric apps, predictive analytics is beginning to enable development of software that is directly used by a consumer. We believe that the growth of consumer-oriented analytic applications will continue and create many entrepreneurial opportunities for the readers of this book.

One key concern in employing these technologies is the loss of privacy. If someone can track the movement of a cell phone, the privacy of that customer is a big issue. Some of the app developers claim that they only need to gather aggregate flow information, not individually identifiable information. But many stories appear in the media that highlight violations of this general principle. Both users and developers of such apps have to be very aware of the deleterious effect of giving out private information as well as collecting such information. We discuss this issue a bit further in Chapter 14.

SECTION 9.10 REVIEW QUESTIONS

1. How does traditional analytics make use of location-based data?
2. How can geocoded locations assist in better decision making?
3. What is the value provided by geospatial analytics?
4. Explore the use of geospatial analytics further by investigating its use across various sectors like government census tracking, consumer marketing, and so forth.
5. Search online for other applications of consumer-oriented analytical applications.
6. How can location-based analytics help individual consumers?
7. Explore more transportation applications that may employ location-based analytics.
8. What other applications can you imagine if you were able to access cell phone location data?

Chapter Highlights

- Big Data means different things to people with different backgrounds and interests.
- Big Data exceeds the reach of commonly used hardware environments and/or capabilities of software tools to capture, manage, and process it within a tolerable time span.
- Big Data is typically defined by three "V"s: volume, variety, velocity.
- MapReduce is a technique to distribute the processing of very large multistructured data files across a large cluster of machines.
- Hadoop is an open source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data.
- Hive is a Hadoop-based data warehousing–like framework originally developed by Facebook.
- Pig is a Hadoop-based query language developed by Yahoo!
- NoSQL, which stands for Not Only SQL, is a new paradigm to store and process large volumes of unstructured, semistructured, and multistructured data.
- Big Data and data warehouses are complementary (not competing) analytics technologies.
- As a relatively new area, the Big Data vendor landscape is developing very rapidly.
- Stream analytics is a term commonly used for extracting actionable information from continuously flowing/streaming data sources.
- Perpetual analytics evaluates every incoming observation against all prior observations.
- Critical event processing is a method of capturing, tracking, and analyzing streams of data to detect certain events (out of normal happenings) that are worthy of the effort.
- Data stream mining, as an enabling technology for stream analytics, is the process of extracting novel patterns and knowledge structures from continuous, rapid data records.
- Cloud computing offers the possibility of using software, hardware, platforms, and infrastructure, all on a service-subscription basis. Cloud computing enables a more scalable investment on the part of a user.

- Cloud-computing–based analytic services offer organizations the latest technologies without significant up-front investment.
- Geospatial data can enhance analytics applications by incorporating location information.
- Real-time location information of users can be mined to develop promotion campaigns that are targeted at a specific user in real time.
- Location information from mobile phones can be used to create profiles of user behavior and movement. Such location information can enable users to find other people with similar interests and advertisers to customize their promotions.
- Location-based analytics can also benefit consumers directly rather than just businesses. Mobile apps are being developed to enable such innovative analytics applications.

Key Terms

Big Data
Big Data analytics
cloud computing
critical event processing
data scientists
data stream mining
geographic information systems (GIS)
Hadoop
Hadoop Distributed File System (HDFS)
Hive
MapReduce
NoSQL
perpetual analytics
Pig
Spark
stream analytics

Questions for Discussion

1. What is Big Data? Why is it important? Where does Big Data come from?
2. What do you think the future of Big Data will be? Will it lose its popularity to something else? If so, what will it be?
3. What is Big Data analytics? How does it differ from regular analytics?
4. What are the critical success factors for Big Data analytics?
5. What are the big challenges that one should be mindful of when considering implementation of Big Data analytics?
6. What are the common business problems addressed by Big Data analytics?
7. In the era of Big Data, are we about to witness the end of data warehousing? Why?
8. What are the use cases for Big Data/Hadoop and data warehousing/RDBMS?
9. Is cloud computing "just an old wine in a new bottle?" How is it similar to other initiatives? How is it different?
10. What is stream analytics? How does it differ from regular analytics?
11. What are the most fruitful industries for stream analytics? What is common to those industries?
12. Compared to regular analytics, do you think stream analytics will have more (or less) use cases in the era of Big Data analytics? Why?
13. What are the potential benefits of using geospatial data in analytics? Give examples.
14. What types of new applications can emerge from knowing locations of users in real time? What if you also knew what they have in their shopping cart, for example?
15. How can consumers benefit from using analytics, especially based on location information?
16. "Location-tracking–based profiling is powerful but also poses privacy threats." Comment.
17. Is cloud computing "just an old wine in a new bottle?" How is it similar to other initiatives? How is it different?
18. Discuss the relationship between mobile devices and social networking.

Exercises

Teradata University Network (TUN) and Other Hands-on Exercises

1. Go to **teradatauniversitynetwork.com**, and search for case studies. Read cases and white papers that talk about Big Data analytics. What is the common theme in those case studies?
2. At **teradatauniversitynetwork.com**, find the SAS Visual Analytics white papers, case studies, and hands-on exercises. Carry out the visual analytics exercises on large data sets and prepare a report to discuss your findings.
3. At **teradatauniversitynetwork.com**, go to the Sports Analytics page. Find applications of Big Data in sports. Summarize your findings.

4. Go to **teradatauniversitynetwork.com**, and search for BSI Videos that talk about Big Data. Review these BSI videos, and answer the case questions related to them.
5. Go to the **teradata.com** and/or **asterdata.com** Web sites. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
6. Go to **IBM.com**. Find at least three customer case studies on Big Data, and write a report where you discuss the commonalities and differences of these cases.
7. Go to **claudera.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
8. Go to **mapr.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
9. Go to **hortonworks.com**. Find at least three customer case studies on Hadoop implementation, and write a report in which you discuss the commonalities and differences of these cases.
10. Go to **marklogic.com**. Find at least three customer case studies on Hadoop implementation, and write a report where you discuss the commonalities and differences of these cases.
11. Go to **youtube.com**. Search for videos on Big Data computing. Watch at least two. Summarize your findings.
12. Go to **google.com**/scholar, and search for articles on stream analytics. Find at least three related articles. Read and summarize your findings.
13. Enter **google.com**/scholar, and search for articles on data stream mining. Find at least three related articles. Read and summarize your findings.
14. Enter **google.com**/scholar, and search for articles that talk about Big Data versus data warehousing. Find at least five articles. Read and summarize your findings.
15. Location-tracking–based clustering provides the potential for personalized services but challenges for privacy. Divide the class into two parts to argue for and against such applications.
16. Enter **YouTube.com**. Search for videos on cloud computing, and watch at least two. Summarize your findings.
17. Enter **Pandora.com**. Find out how you can create and share music with friends. Explore how the site analyzes user preferences.
18. Enter **Humanyze.com**. Review various case studies and summarize one interesting application of sensors in understanding social exchanges in organizations.
19. The objective of the exercise is to familiarize you with the capabilities of smartphones to identify human activity. The data set is available at **archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones**.

    It contains accelerometer and gyroscope readings on 30 subjects who had the smartphone on their waist. The data is available in a raw format and involves some data preparation efforts. Your objective is to identify and classify these readings into activities like walking, running, climbing, and such. More information on the data set is available on the download page. You may use clustering for initial exploration and to gain an understanding of the data. You may use tools like R to prepare and analyze this data.

References

Adapted from **Alteryx.com**. Great Clips. **alteryx.com/sites/default/files/resources/files/case-study-great-chips.pdf** (accessed September 2018).

Adapted from **Snowflake.net**. (n.d.). "Chime Delivers Personalized Customer Experience Using Chime." **www.snowflake.net/product** (accessed September 2018).

Adshead, A. (2014). "Data Set to Grow 10-fold by 2020 as Internet of Things Takes Off." **www.computerweekly.com/news/2240217788/Data-set-to-grow-10-fold-by-2020-as-internet-of-things-takes-off** (accessed September 2018).

Altaweel, Mark. "Accessing Real-Time Satellite Imagery and Data." GIS Lounge, 1 Aug. 2018, **www.gislounge.com/accessing-real-time-satellite-imagery/**.

Amodio, M. (2015). Salesforce adds predictive analytics to Marketing Cloud. Cloud Contact **Center.cloudcontactcenterzone.com/topics/cloud-contact-center/articles/413611-salesforce-adds-predictive-analytics-marketing-cloud.htm** (accessed September 2018).

Asamoah, D., & R. Sharda. (2015). Adapting CRISP-DM process for social network analytics: Application to healthcare. *In AMCIS 2015 Proceedings*. **aisel.aisnet.org/amcis2015/BizAnalytics/GeneralPresentations/33/** (accessed September 2018).

Asamoah, D., R. Sharda, A. Zadeh, & P. Kalgotra. (2016). "Preparing a Big Data Analytics Professional: A Pedagogic Experience." In *DSI 2016 Conference*, Austin, TX.

Awadallah, A., & D. Graham. (2012). "Hadoop and the Data Warehouse: When to Use Which." **teradata.com/white-papers/Hadoop-and-the-Data-Warehouse-When-to-Use-Which** (accessed September 2018).

Blogs.ESRI.com. "From Customers to CxOs, Starbucks Delivers World-Class Service." (2014). **https://blogs.esri.com/esri/ucinsider/2014/07/29/starbucks/** (accessed September 2018).

Broniatowski, D. A., M. J. Paul, & M. Dredze. (2013). "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012–2013 Influenza Epidemic." *PloS One, 8*(12), e83672.

Cisco. (2016). "The Zettabyte Era: Trends and Analysis." **cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf** (accessed October 2018).

ComputerWeekly.com. (2016). "Big-Data and Open Source Cloud Technology Help Gulf Air Pin Down Customer Sentiment." **www.computerweekly.com/news/450297404/Big-data-and-open-source-cloud-technology-help-Gulf-Air-pin-down-customer-sentiment** (accessed September 2018).

CxOtoday.com. (2014). "Cloud Platform to Help Pharma Co Accelerate Growth." **www.cxotoday.com/story/mankind-pharma-to-drive-growth-with-softlayers-cloud-platform/** (accessed September 2018).

Dalininaa, R., "Using Natural Language Processing to Analyze Customer Feedback in Hotel Reviews," **awww.datascience.com/resources/notebooks/data-science-summarize-hotel-reviews** (Accessed October 2018).

DataStax. "Customer Case Studies." **datastax.com/resources/casestudies/eBay** (accessed September 2018).

Davis, J. (2015). "Salesforce Adds New Predictive Analytics to Marketing Cloud. Information Week." **informationweek.com/big-data/big-data-analytics/salesforce-adds-new-predictive-analytics-to-marketing-cloud/d/d-id/1323201** (accessed September 2018).

Dean, J., & S. Ghemawat. (2004). "MapReduce: Simplified Data Processing on Large Clusters." **research.google.com/archive/mapreduce.html** (accessed September 2018).

Delen, D., M. Kletke, & J. Kim. (2005). "A Scalable Classification Algorithm for Very Large Datasets." *Journal of Information and Knowledge Management, 4*(2), 83–94.

Demirkan, H., & D. Delen. (2013, April). "Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud." *Decision Support Systems, 55*(1), 412–421.

Digit.HBS.org. (2015). "Starbucks: Brewing up a Data Storm!" **https://digit.hbs.org/submission/starbucks-brewing-up-a-data-storm/** (accessed September 2018).

Dillow, C. (2016). "What Happens When You Combine Artificial Intelligence and Satellite Imagery." **fortune.com/2016/03/30/facebook-ai-satellite-imagery/** (accessed September 2018).

Ekster, G. (2015). "Driving Investment Performance with Alternative Data." **integrity-research.com/wp-content/uploads/2015/11/Driving-Investment-Performance-With-Alternative-Data.pdf** (accessed September 2018).

Henschen, D. (2016). "Salesforce Reboots Wave Analytics, Preps IoT Cloud." *ZD Net*. **zdnet.com/article/salesforce-reboots-wave-analytics-preps-iot-cloud/** (accessed September 2018).

Higginbotham, S. (2012). "As Data Gets Bigger, What Comes after a Yottabyte?" **gigaom.com/2012/10/30/as-data-gets-bigger-what-comes-after-a-yottabyte** (accessed September 2018).

Hope, B. (2015). "Provider of Personal Finance Tools Tracks Bank Cards Sells Data to Investors." *Wall Street Journal*. **wsj.com/articles/provider-of-personal-finance-tools-tracks-bank-cards-sells-data-to-investors-1438914620** (accessed September 2018).

Jonas, J. (2007). "Streaming Analytics vs. Perpetual Analytics (Advantages of Windowless Thinking)." **jeffjonas.typepad.com/jeff_jonas/2007/04/streaming_analy.html** (accessed September 2018).

Kalgotra, P., & R. Sharda. (2016). "Rural Versus Urban Comorbidity Networks." Working Paper, Center for Health Systems and Innovation, Oklahoma State University.

Kalgotra, P., R. Sharda, & J. M. Croff. (2017). "Examining Health Disparities by Gender: A Multimorbidity Network Analysis of Electronic Medical Record." *International Journal of Medical Informatics, 108*, 22–28.

Kelly, L. (2012). "Big Data: Hadoop, Business Analytics, and Beyond." **wikibon.org/wiki/v/Big_Data:_Hadoop,_Business_Analytics_and_Beyond** (accessed September 2018).

Moran, P. A. (1950). "Notes on Continuous Stochastic Phenomena." *Biometrika*, 17–23.

"Overstock.com: Revolutionizing Data and Analytics to Connect Soulfully with their Customers," at **https://www.teradata.com/Resources/Videos/Overstock-com-Revolutionizing-data-and-analy** (accessed October 2018).

"Overstock.com Uses Teradata Path Analysis To Boost Its Customer Journey Analytics," March 27, 2018, at **https://www.retailitinsights.com/doc/overstock-com-uses-teradata-path-analysis-boost-customer-journey-analytics-0001** (accessed October 2018).

Palmucci, J., "Using Apache Spark for Massively Parallel NLP," at **http://engineering.tripadvisor.com/using-apache-spark-for-massively-parallel-nlp/** (accessed October 2018).

RedHat.com. (2016). "Gulf Air's Big Data Innovation Delivers Deeper Customer Insight." **https://www.redhat.com/en/success-stories** (accessed September 2018).

RedHat.com. (2016). "Gulf Air Builds Private Cloud for Big Data Innovation with Red Hat Technologies." **https://www.redhat.com/en/about/press-releases/gulf-air-builds-private-cloud-big-data-innovation-red-hat-technologies** (accessed September 2018).

Russom, P. (2013). "Busting 10 Myths about Hadoop: the Big Data Explosion." TDWI's *Best of Business Intelligence, 10*, 45–46.

Sarasohn-Kahn, J. (2008). *The Wisdom of Patients: Health Care Meets Online Social Media*. Oakland, CA: California HealthCare Foundation.

Shaw, C. (2016). "Satellite Companies Moving Markets." **quandl.com/blog/alternative-data-satellite-companies** (accessed September 2018).

Steiner, C. (2009). "Sky High Tips for Crop Traders" (accessed September 2018).

St Louis, C., & G. Zorlu. (2012). "Can Twitter Predict Disease Outbreaks?" *BMJ*, 344.

Tableau white paper. (2012). "7 Tips to Succeed with Big Data in 2013." **cdnlarge.tableausoftware.com/sites/default/files/whitepapers/7-tips-to-succeed-with-big-data-in-2013.pdf** (accessed September 2018).

Tartar, Andre, et al. "All the Things Satellites Can Now See From Space." **Bloomberg.com**, Bloomberg, 26 July 2018, **www.bloomberg.com/news/features/2018-07-26/all-the-things-satellites-can-now-see-from-space** (accessed October 2018).

Thusoo, A., Z. Shao, & S. Anthony. (2010). "Data Warehousing and Analytics Infrastructure at Facebook." In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* (p. 1013).

Turner, M. (2015). "This Is the Future of Investing, and You Probably Can't Afford It." **businessinsider.com/hedge-funds-are-analysing-data-to-get-an-edge-2015-8** (accessed September 2018).

Watson, H. (2012). "The Requirements for Being an Analytics-Based Organization." *Business Intelligence Journal, 17*(2), 42–44.

Watson, H., R. Sharda, & D. Schrader. (2012). "Big Data and How to Teach It." *Workshop at AMCIS*, Seattle, WA.

Wheeler, C. (2014). "Going Big with GIS." **www.esri.com/esri-news/arcwatch/0814/going-big-with-gis** (accessed October 2018).

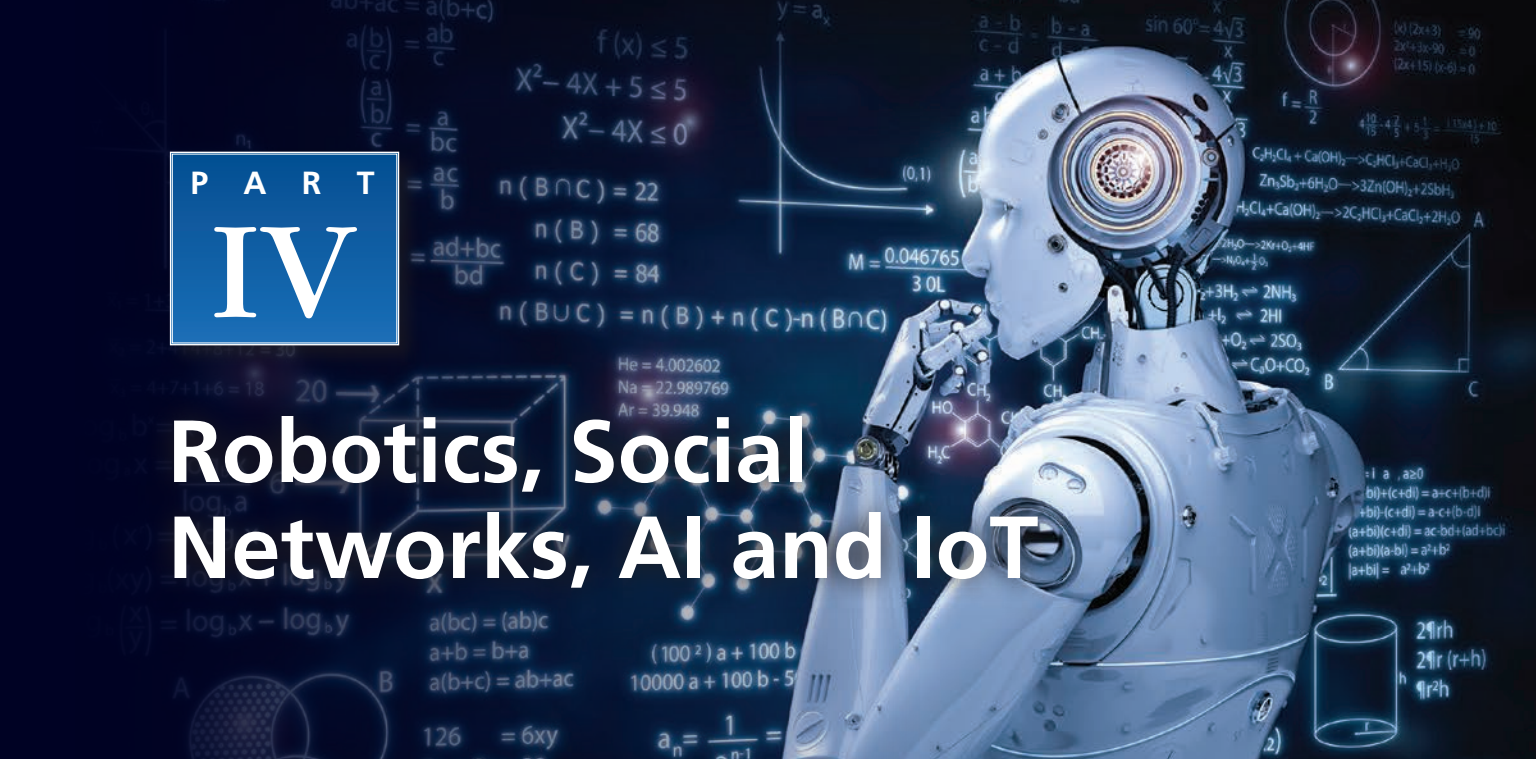
White, C. (2012). "MapReduce and the Data Scientist." Teradata Vantage White Paper. **teradata.com/white-paper/MapReduce-and-the-Data-Scientist** (accessed September 2018).

Wikipedia.com. "Petabyte." en.wikipedia.org/wiki/Petabyte (accessed September 2018).

Zadeh, A. H., H. M. Zolbanin, R. Sharda, & D. Delen. (2015). "Social Media for Nowcasting the Flu Activity: Spatial-Temporal and Text Analysis." *Business Analytics Congress, Pre-ICIS Conference*, Fort Worth, TX.

Zikopoulos, P., D. DeRoos, K. Parasuraman, T. Deutsch, D. Corrigan, & J. Giles. (2013). *Harness the Power of Big Data*. New York: McGraw-Hill.

"Zion China uses Azure IoT, Stream Analytics, and Machine Learning to Evolve Its Intelligent Diabetes Management Solution," **www.codeproject.com/Articles/1194824/Zion-China-uses-Azure-IoT-Stream-Analytics-and-M** (accessed October 2018) and **https://microsoft.github.io/techcasestudies/iot/2016/12/02/IoT-ZionChina.html** (accessed October 2018).



PART IV

Robotics, Social Networks, AI and IoT

CHAPTER

10

Robotics: Industrial and Consumer Applications

LEARNING OBJECTIVES

- Discuss the general history of automation and robots
- Discuss the applications of robots in various industries
- Differentiate between industrial and consumer applications of robots
- Identify common components of robots
- Discuss impacts of robots on future jobs
- Identify legal issues related to robotics

Chapter 2 briefly introduced robotics, an early and practical application of concepts developed in AI. In this chapter, we present a number of applications of robots in industrial as well as personal settings. Besides learning about the already deployed and emerging applications, we identify the general components of a robot. In the spirit of managerial considerations, we also discuss the impact of robotics on jobs as well as related legal issues. Some of the coverage is broad and impacts all other artificial intelligence (AI), so it may seem to overlap a bit with Chapter 14. But the focus in this chapter is on physical robots, not just software-driven applications of AI.

This chapter has the following sections:

10.1 OPENING VIGNETTE: Robots Provide Emotional Support to Patients and Children

As discussed in this chapter, robots have impacted industrial manufacturing and other physical activities. Now, with the research and evolution of AI, robotics can straddle the social world. For example, hospitals today make an effort to give social and emotional support to patients and their families. This support is especially sensitive when offering treatment to children. Children in a hospital are in an unfamiliar environment with medical instruments attached to them, and in many cases, doctors may recommend movement restrictions. This restriction leads to stress, anxiety, and depression in children and consequently in their family members. Hospitals try to provide childcare support specialist or companion pet therapies to reduce the trauma. These therapies prepare children and their parents for future treatment and provide them with temporary emotional support with their interactions. Due to the small number of such specialists, there is a gap between demand and supply for childcare specialists. Also, it is not possible to provide pet therapy at many centers due to the fear of allergies, dust, and bites that may cause the patient's condition to be aggravated. To fill these gaps, the use of social robots is being explored to resolve depression and anxiety among children. A study (Jeong et al., 2015) found that the physical presence of a robot is more effective concerning emotional response as compared to a virtual machine interaction in a pediatric hospital center.

Researchers have known for a long time (e.g., Goris et al., 2010) that more than 60 percent of human communication is not verbal but rather occurs through facial expressions. Thus, a social robot has to be able to provide emotional communication like a child specialist. One popular robot that is providing such support is Huggable. With the help of AI, Huggable is equipped to understand facial expressions, temperament, gestures, and human cleverness. It is like a staff member added to the team of specialists who provide children some general emotional health assistance.

Huggable looks like a teddy bear having a ringed arrangement. A furry soft body provides a childish look to it and hence is perceived as a friend by the children. With its mechanical arms, Huggable can perform specific actions quickly. Rather than sporting high-tech devices, a Huggable robot is composed of an Android device whose microphone, speaker, and camera are in its internal sensors, and a mobile phone that acts as the central nervous system. The Android device enables the communication between the internal sensors and teleoperation interface. Its segmental arm components enable an easy replacement of sensors and hence increase its reusability. These haptic sensors along with AI enable it to process the physical touch and use it expressively.

Sensors incorporated in a Huggable transmit physical touch and pressure data to the teleoperation device or external device via an IOIO board. The Android device receives the data from the external sensors and transmits them to the motors that are attached to the body of the robot. These motors enable the movement of the robot. The capacitors are placed at various parts of the robot, known as *pressure points*. These pressure points enable the robot to understand the pain of a child who is unable to express it verbally but may be able to touch the robot to convey the pain. The Android device interprets the physical touch and pressure sensor data in a meaningful way and responds effectively. The Android phone enables communication between the other devices while keeping the design minimalistic. The computing power of the robot and the Android device is good enough to allow real-time communication with a child. Figure 10.1 exhibits a schematic of the Huggable robot.

Huggable has been used with children undergoing treatment at Boston Children's Hospital. Reportedly, Aurora, a 10-year-old who had leukemia, was being treated at Dana-Farber/Boston Children's Cancer and Blood Disorder Center. According to Aurora's parents, "There were many activities to do at the hospital but the Huggable being there

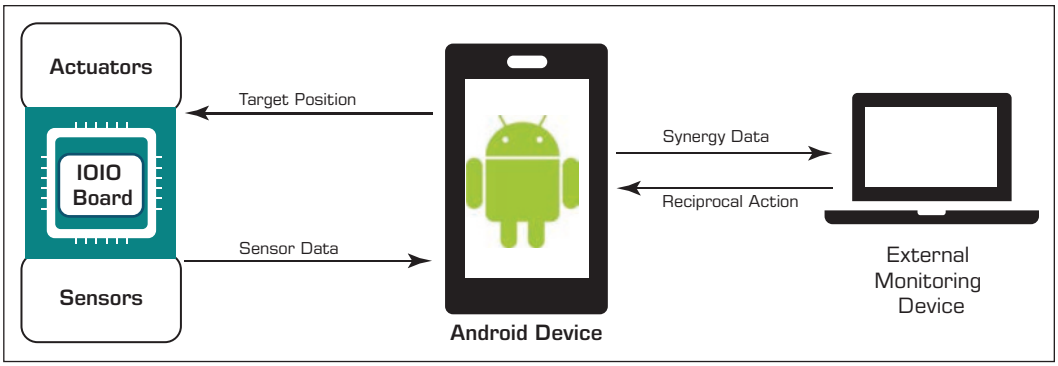


FIGURE 10.1 **A General Schematic of a Huggable Robot.**

is great for kids." Beatrice, another child who visits the hospital frequently due to her chronic condition, misses her classes and friends and is unable to do anything that a typical child of her age would do. She was nervous and disliked the process of treatment, but during her interplay with the Huggable, she was more willing to take medicine as if it were the most natural activity to do. She recommended the robot to be a bit faster so that the next time she could play peek-a-boo correctly.

During these interactions with Huggable, children were seen hugging it, holding its hand, tickling it, giving it high-fives, and treating it as someone they need for support. Children were polite with it and used expressions such as "no, thank you" and "one second, please." In the end, when bidding it goodbye, one child hugged the Huggable, and another wished to play with it longer.

Another benefit of such emotional support robots is in the prevention of infections. Patients may have contagious diseases, but the robots are sterilized after each use to prevent infection from spreading. Thus, Huggable not only provides support to children but also can be a useful tool for reducing the spread of infectious diseases.

A recently reported study by researchers at MIT's Media Lab highlighted the differences between social robots such as Huggable and other virtual interaction technologies. A group of 54 children who were in a hospital were given three distinct social interactions: a cute regular teddy bear, a virtual persona of Huggable on a tablet, and a social robot. The bear offered a physical model but not social dialogue. A virtual version of Huggable on the tablet provided linguistic engagement, conversed with the humans in the same way, and possessed the same features as the robot but was a 3D virtual version of the Huggable robot. Both the virtual character and robot were operated by a teleoperator, and hence they perceived the interaction and responded in the same manner. Children were given one of the three interactions to play with based in groups according to age and gender. Necessary information was provided to the children by the care specialist, and the virtual character and robot were handled separately by these specialists just outside the room. IBM Watson's tone Analyzer was used to attempt to identify five human emotions and five personality traits. Interactions with each of these three types of virtual agents by children were videotaped and analyzed by the researchers. The results of this experiment were quite interesting. These results showed that the children gazed more at the virtual character and the robot as compared to the bear. Touches between the children and the virtual agents were the highest with the Huggable robot followed by the virtual character Huggable and the bear. Also, the children took care of the Huggable robot and did not push or pull it. Interestingly, a few children responded to the virtual character on the tablet violently even when it made ouch sounds. The poor teddy bear was thrown and kicked around playfully. These results show that the children connected with the robotic Huggable more than the other two options.

SOCIAL ROBOT FOR OLDER ADULTS: PARO

Major countries in the world will soon have population rates of people aged 65 and older exceed that of the younger population by 2050. The emotional support older adults need cannot be ignored where geographical separation and the technological divide have made it difficult for them to connect with their families. Paro, a social robot, is designed to interact with humans and is marketed as a robot used for older adults at nursing homes. Paro mainly acts as pet therapy; it can also be immensely useful where pet therapy becomes inadvisable in hospitals due to the risk of infections.

Paro interprets the human touch, and it can also capture limited speech, express a limited set of vocal utterances, and move its head. Paro is not a mobile robot and resembles a seal. It was tested at two nursing homes (Broekens et al., 2009) with 23 patients. The results showed that social robots like Paro increase the social interaction. Paro not only brought smiles to patients' faces but also some vivid, happy experiences to occupants. Even though Paro did not provide a complete response that humans do, many patients found the responses meaningful and connected emotionally to it. These robots can help break the monotonous routine of older adults and add some joy to their lives. It provides them with a feeling of being wanted and self-esteem, lowering stress and anxiety levels.

QUESTIONS FOR THE OPENING VIGNETTE

1. What characteristics would you expect to have in a robot that provides emotional support to patients?
2. Can you think of other applications where robots such as the Huggable can play a helpful role?
3. Visit the website **https://www.universal-robots.com/case-stories/aurolab/** to learn about collaborative robots. How could such robots be useful in other settings?

WHAT WE CAN LEARN FROM THIS VIGNETTE

As we have seen in various chapters throughout this book, AI is opening many interesting and unique applications. The stories about the Huggable and Paro introduce us to the idea of using robots for one of the most difficult aspects of work – to provide emotional support to patients, both children and adults. Combinations of technologies such as machine learning, voice synthesis, voice recognition, natural language processing, machine vision, automation, micromachines, and so on make it possible to combine these technologies to satisfy many needs. The applications can come entirely in virtual forms such as IBM Watson, which won the *Jeopardy!* game implementing industrial automation, producing self-driving cars, and even providing emotional support as noted in this opening vignette. We will see many examples of similar applications in this chapter.

Sources: J. Broekens, M. Heerink, & H. Rosendal. (2009). "Assistive Social Robots in Elderly Care: A Review." *Gerontechnology, 8*, pp. 94–103. doi: 10.4017/gt.2009.08.02.002.00; S. Fallon. (2015). "A Blue Robotic Bear to Make Sick Kids Feel Less Blue." **https://www.wired.com/2015/03/blue-robotic-bear-make-sick-kids-feel-less-blue/** (accessed August 2018). Also see the YouTube video at **https://youtu.be/UaRCCA2rRR0** (accessed August 2018); K. Goris et al. (2010, September). "Mechanical Design of the Huggable Robot Probo." Robotics & Multibody Mechanics Research Group. Brussels, Belgium: Vrije Universiteit Brussel; S. Jeong et al. (2015). "A Social Robot to Mitigate Stress, Anxiety, and Pain in Hospital Pediatric Care." *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*; S. Jeong & D. Logan. (2018, April 21–26). "Huggable: The Impact of Embodiment on Promoting Socio-emotional Interactions for Young Pediatric Surgeons." MIT Media Lab, Cambridge, MA, CHI 2018, Montréal, Quebec, Canada.

10.2 OVERVIEW OF ROBOTICS

Every robotics scientist has her or his own view about the definition of robot. But a common notion of robot is a machine or a physical device or software that with the cooperation of AI can accomplish a responsibility autonomously. A robot can sense and affect the environment. Applications of robotics in our day-to-day lives have been increasing. This evolution and use of technologies are called the *fourth industrial revolution*. Applications of robotics in manufacturing, health, and information technology (IT) fields in the last decade have led to rapid development in changing the future of industries. Robots are moving from just performing preselected repetitive tasks (**automation**) and being unable to react to unforeseen circumstances (Ayres and Miller, 1981) to performing specialized tasks in healthcare, manufacturing, sports, financial services – virtually every industry. This capability of adaptation to new situations leads to **autonomy**, a sea change from previous generations of robots. Chapter 2 introduced a definition of robots and provided some applications in selected industries. In this chapter, we will supplement that introduction with various applications and take a slightly deeper dive into the topic.

Although our imagination of a robot may be based on the R2D2 or C3-PO from the *Star Wars* movies, we have experienced robots in many other ways. Factories have been using robots for a long time (see Section 10.3) for manufacturing. On the consumer side, an early application was Roomba, a robot that can clean floors on its own. Perhaps the best example of robots that we will all experience soon if not already is an autonomous (self-driving) car. *Tech Republic* called the self-driving car the first robot we will all learn to trust. We will dig a bit deeper into self-driving vehicles in Section 10.7. With the growth in machine learning, especially image recognition systems, applications of robots are increasing in virtually every industry. Robots can cut sausages into the right size pieces for pizza and can automatically determine that the right number and type of peperoni pieces have been placed on a pizza before it is baked. Surgeries conducted by and with the assistance of robots are growing at a rapid pace. Section 10.4 provides many illustrative applications of robots. Then Section 10.7 discusses self-driving cars as another category of robots.

SECTION 10.2 REVIEW QUESTIONS

1. Define *robot*.
2. What is the difference between automation and autonomy?
3. Give examples of robots in use. Find recent applications online and share with the class.

10.3 HISTORY OF ROBOTICS

Wikipedia includes an interesting history of robotics. Humans have been fascinated with the idea of machines serving us for a long time. The first idea of robotics was conceptualized in 320 BC when Aristotle, a Greek philosopher, stated, "If every tool, when ordered, or even of its own accord, could do the work that befits it, then there would be no need either of apprentices for the master workers or of slaves for the lords." In 1495, Leonardo Da Vinci drafted strategies and images for a robot that looked like a human. Between 1700 and 1900, various automatons were created, including an excellent automation structure built by Jacques De Vaucanson, who made one clockwork duck that could flap its wings, quack, and appear to eat and digest food.

Throughout the industrial revolution, robotics was triggered by the advances in steam power and electricity. As consumer demand increased, engineers strove to devise new methods to increase production by automation and create machines that can perform

the tasks that were dangerous for a human to do. In 1893, "Steam Man," a prototype for a humanoid robot, was proposed by Canadian professor George Moore. It was composed of steel and powered by a steam engine. It could walk autonomously at nearly nine miles per hour and could even pull relatively light loads. In 1898, Nikola Tesla exhibited a submarine prototype. These events led to the integration of robotics in manufacturing, space, defense, aerospace, medicine, education, and entertainment industries.

In 1913, the world's first moving conveyor belt assembly line was started by Henry Ford. With the aid of a conveyor belt, a car could be assembled in 93 minutes. Later in 1920, the term *robot* was coined by Karel Capek in his play *Rossum's Universal Robots*. Then a toy robot, Lilliput, was manufactured in Japan.

By the 1950s, innovators were creating machines that could handle dangerous, repetitive tasks for defense and industrial manufacturing. Since the robots were primarily designed for heavy-duty industries, they were required to pull, lift, move, and push the same way humans did. Thus, many robots were designed like a human arm. Examples include a spray-painting gadget for a position-controlling apparatus by W. L. V. Pollard in 1938. DeVilbiss Company acquired this robot and later became a leading supplier of the robotic arms in the United States.

In the mid-1950s, the first commercial robotic arm, Planetbot, was developed, and General Motors later used it in a manufacturing plant for the production of radiators. A total of eight Planetbots were sold. According to the company, it could perform nearly 25 movements and could be reset in minutes to perform another set of operations. However, Planetbot did not achieve the desired results due to the unusual behavior of the hydraulic fuel inside it.

George Devol and Joe Engelberger designed Unimate to automate the manufacturing of TV picture tubes. It weighed close to 4,000 pounds and was controlled by preprogrammed commands fed on a magnetic drum. Later this was used by General Motors Corporation for production to sequence and stack hot die-cast metal components. This arm with specific upgrades became one of the famous features in assembly lines. A total of 8,500 machines was sold, and half of them went to the automotive industries. Later Unimate was modified to perform spot welding, die casting, and machine tool stacking.

In the 1960s, Ralph Mosher and his team created two remotely operated robotic arms, Handyman and Man-mate. A Handyman was a two-arm electro-hydraulic robot, and the design of the Man-mate's arm was based on the human spine. The arms gave the robots the flexibility for artifact examination procedures. The fingers were designed in a way that they could grasp objects via a single command.

New mobile robots came into the picture. The first one, Shakey, was developed in 1963. It could move freely, avoiding obstacles in its path. A radio antenna was attached to its head. It had a vision sensor atop a central processing unit. Shakey was attached to two wheels, and its two sensors could sense obstacles. Using logic-based problem solving, it could recognize the shape of objects, move them, or go around them.

The space race started by Russia's Sputnik and embraced by the United States led to many technology advances leading to growth of robotics. In 1976, during NASA's mission to Mars, a Viking lander was created for the atmospheric conditions of Mars. Its arms opened out and created a tube to gather samples from the Mars surface. There were some technical issues during the mission, but the scientists were able to fix them remotely.

In 1986, the first LEGO-based educational products were put on the market by Honda. In 1994, Dante II, an eight-legged walking robot built by Carnegie Mellon University, collected the volcanic gas sample from Mount Spur.

Robotics expanded exponentially as more research and money were invested. Robotic applications and research spread to Japan, Korea, and European nations. It is estimated that by 2019, there will be close to 2.6 million significant robots. Robots have applications in the fields of social support, defense, toys and entertainment, healthcare,

food, and rescue. Many robots are now moving into next stages, going from deep-sea to interplanetary and extrasolar research. And as noted, self-driving cars will bring robots to the masses. We review several robot applications in the following sections.

SECTION 10.3 REVIEW QUESTIONS

1. Identify some of the key milestones in the history of manufacturing that have led to the current interest in robotics.
2. How would Shakey's capabilities compare to today's robots?
3. How have robots helped with space missions?

10.4 ILLUSTRATIVE APPLICATIONS OF ROBOTICS

This section highlights examples of robot applications in various industries. Each of these is presented as a mini-application case, with the discussion questions presented at the end of the section.

Changing Precision Technology

A mobile production company in China, Changing Precision Technology switched to the use of robotic arms to produce parts for mobile phones. The company previously employed 650 workers to operate the factory. Now, robots perform most of its operations, and the company has reduced its workforce to 60, decreasing the human workforce by 90 percent. In the future, the company intends to drop its employee count to about 20. With the robots in place, the company not only has achieved an increase in production of 250 percent but also cut the defect levels from 25 percent to a mere 5 percent.

Compiled from C. Forrest. (2015). "Chinese Factory Replaces 90% of Humans with Robots, Production Soars." TechRepublic. **https://www.techrepublic.com/article/chinese-factory-replaces-90-of-humans-with-robots-production-soars/** (accessed September 2018); J. Javelosa & K. Houser. (2017). "Production Soars for Chinese Factory Who Replaced 90% of Employees with Robots." Future Society. **https://futurism.com/2-production-soars-for-chinese-factory-who-replaced-90-of-employees-with-robots/** (accessed September 2018).

Adidas

Adidas is a worldwide leading sportswear manufacturer. Keeping trends, innovation, and customization in mind, Adidas has started to automate factories such as Speedfactory in Ansbach, Germany, and Atlanta, Georgia. A conventional supply chain from the raw materials to final product takes around two months, but with automation, it takes just a few days or weeks. The implementation of robotics there was different from that of other manufacturing industries because the raw materials used in shoes manufactured by Adidas are soft textile materials. Adidas is working with the company Oechsler to implement the robotics in its supply chain. Adidas uses technologies such as additive manufacturing, robotic arms, and computerized knitting. At the Speedfactory, the robot that makes a part of a sneaker attaches a scannable QR code to the part. During quality check, if any part of the product turns out to be faulty, the robot that created it is thus traceable and repaired. Adidas has optimized this process, which offers the company the option to roll out a few thousands of customized shoes in the market and see how it performs and optimize the process accordingly. In the next few years, the company plans to roll out around 1 million pairs of the custom styles annually. In the long term, this strategy supports moving from manufacturing large stocks of inventory to creating the products on demand.

Compiled from "Adidas's High-Tech Factory Brings Production Back to Germany." (2017, January 14). *The Economist*. **https://www.economist.com/business/2017/01/14/adidass-high-tech-factory-brings-production-back-to-germany** (accessed September 2018); D. Green. (2018). "Adidas Just Opened a Futuristic New Factory – and It Will Dramatically Change How Shoes Are Sold." Business Insider. **http://www.businessinsider.com/adidas-high-tech-speedfactory-begins-production-2018-4** (accessed September 2018).

BMW Employs Collaborative Robots

The increased use of AI and automation in industries has resulted in the development of robots. Yet, human cognitive capabilities are irreplaceable. The combination of robots and humans has been achieved using collaborative robots at a BMW manufacturing unit. By doing so, the company has maximized the efficiency of its production unit and modernized the work environment.

BMW's Spartanburg, South Carolina, plant has employed 60 collaborative robots that work side by side with its human workforce. These robots, for example, furnish the interior of BMW car doors with sound and moisture insulation. This sealing protects the electronic equipment that is fixed on the door and the vehicle as a whole from moisture. Previously, human workers performed this intensive task of fixing the foil with the adhesive beads by using a manual roller. With the use of cobots, a robot's arms perform this task with precision. Cobots run on low speed and stop immediately as soon as the sensors detect any obstacle in their way to maintain the safety of assembly-line workers.

At BMW's Dingolfing factory located in Germany, a lightweight cobot is ceiling mounted in the axle transmission assembly area to pick up bevel gears. These gears can weigh up to 5.5 kilos. The cobot fits the bevel gears accurately, avoiding damage to the gear wheels.

Compiled from M. Allinson. (2017, March 4). "BMW Shows Off Its Smart Factory Technologies at Its Plants Worldwide." BMW Press Release. Robotics and Automation. **https://roboticsandautomationnews.com/2017/03/04/bmw-shows-off-its-smart-factory-technologies-at-its-plants-worldwide/11696/** (accessed September 2018); "Innovative Human-Robot Cooperation in BMW Group Production." (2013, October 9). **https://www.press.bmwgroup.com/global/article/detail/T0209722EN/innovative-human-robot-cooperation-in-bmw-group-production?language=en** (accessed September 2018).

Tega

Tega is a social bot intended to provide extended support to preschoolers by engaging them via storytelling and offering help with vocabulary. Like Huggable, Tega is an Android-based robot and resembles an animation character. It has an external camera and onboard speakers and is designed to run for up to six hours before needing a recharge. Tega uses Android capabilities for expressive eyes, computation abilities, and physical movements. Children's response is fed to the Tega as a reward signal into a reinforcement learning algorithm. Tega uses a social controller, sensor processing, and motor control for moving its body and tilting and rotating left or right.

Tega is designed not only to tell stories but also to hold a conversation about the stories. With the help of an app on a tablet, Tega interacts with a child as a peer and teammate, not as an educator. Children communicate with the tablet, and Tega provides the feedback and reactions by watching the children's emotional states. Tega also offers help with vocabulary and understands a child's physical and emotional responses, enabling it to build a relationship with the child. The tests have shown that Tega can positively impact a child's interest in education, free thinking, and mental development. For more information, watch the video at **https://www.youtube.com/watch?v=16in922JTsw**.

Compiled from E. Ackerman. (2016). *IEEE Spectrum*. **http://spectrum.ieee.org/automaton/robotics/home-robots/tega-mit-latest-friendly-squishable-social-robot** (March 5, 2017); J. K. Westlund et al. (2016). "Tega: A Social Robot." Video Presentation. *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction*; H. W. Park et al. (2017). "Growing Growth Mindset with a Social Robot Peer." *Proceedings of the Twelfth ACM/IEEE International Conference on Human Robot Interaction*; Personal Robots Group. (2016). **https://www.youtube.com/watch?v=sF0tRCqvyT0** (accessed September 2018); Personal Robots Group, MIT Media Lab. (2016). *AAAS*. **https://www.eurekalert.org/pub_releases/2016-03/nsf-rlc031116.php** (accessed September 2018).

San Francisco Burger Eatery

Flipping burgers is considered a low-pay, mundane task that provides many people with employment at a low salary. Such jobs are likely to disappear over time because of robots. One such implementation of robotics in the food industry is at a burger restaurant in San Francisco. The burger-making machine is not a traditional robot sporting arms and legs that can move around and work as a human. Instead, it is a complete burger prep device that can work from prepping a burger for cooking and bringing together a full meal. It blends the robotic power in bringing the right taste with the help of a Michelin-star chef's recipes and being friendly on the pocket. The restaurant has put in place two 14-foot-long machines that can make around 120 burgers per hour. Each machine has 350 sensors, 20 computers, and close to 7,000 parts.

Buns, onions, tomatoes, pickles, seasoning, and sauces are filled in transparent tubes over a conveyor belt. Once an order is placed via a mobile device, it takes close to five minutes to prepare the order. First, air pressure pushes a burger brioche roll from the transparent tube on the conveyor belt. Different components of the robot work one after the other to prepare the order, from slicing the roll in two halves, applying butter on the bun, shredding vegetables, and dropping the sauces. Also, a light specialized grip is placed on the patty to keep it intact and to bake it per the recipe. With the use of thermal sensors and an algorithm, the cooking time and temperature of the patty are determined, and once cooked, the patty is placed on the bun by a robotic arm. Workers receive a notification via an Apple watch when there is an issue with the machine regarding a malfunction on an order or the need for refills on supplies.

Compiled from "A Robot Cooks Burgers at Startup Restaurant Creator." (2018). TechCrunch. **https://techcrunch.com/video/a-robot-cooks-burgers-at-startup-restaurant-creator/** (accessed September 2018); L. Zimberoff. (2018, June 21). "A Burger Joint Where Robots Make Your Food." **https://www.wsj.com/articles/a-burger-joint-where-robots-make-your-food-1529599213** (accessed September 2018).

Spyce

Using robots to make affordable foods is demonstrated by a fast-food restaurant operating in Boston that serves grain dishes and salad bowls. Spyce is a budget-friendly restaurant founded by MIT engineering graduates. Michael Farid created the robots that can cook. This restaurant employs few people with good pay and employs robots to do much of the fast-food work.

Orders are placed at a kiosk with touch screens. Once the order is confirmed, the mechanized systems start preparing the food. Ingredients are placed in refrigerated bins that are passed via transparent tubes and are collected using a mobile device that delivers the ingredients to the requested pot. A metal plate attached to the side of the robotic pot heats the food. A temperature of about 450 degrees Fahrenheit is maintained, and the food is tumbled for nearly two minutes and cooked. This resembles clothes being washed in a machine. Once the meal is ready, the robotic pot tilts and transfers the food to a bowl. After each cooking round, the robotic pot washes itself with a high-pressure hot water stream and then returns to its initial position, ready to cook the next meal. The customer name is also added to the bowl. The meal is then served by a human after any final changes. Spyce is also trying to put in place a robot that can cook pancakes.

Compiled from B. Coxworth. (2018, May 29). "Restaurant Keeps Its Prices Down – With a Robotic Kitchen." New Atlas. **https://newatlas.com/spyce-restaurant-robotic-kitchen/54818/** (accessed September 2018); J. Engel. (2018, May 3). "*Spyce, MIT-Born Robotic Kitchen Startup, Launches Restaurant: Video.*" Xconomy. **https://www.xconomy.com/boston/2018/05/03/spyce-mit-born-robotic-kitchen-startup-launches-restaurant-video/** (accessed September 2018).

Mahindra & Mahindra Ltd.

As the population increases, the agricultural industry is expanding to keep up with demand. To keep increasing the food supply at a reasonable cost and to maintain quality, the Indian multinational firm Mahindra & Mahindra Ltd. is seeking to improve the process of harvesting tabletop grapes. The company is establishing a research and development center at Virginia Polytechnic Institute and University. It will work with other Mahindra centers situated in Finland, India, and Japan.

The grapes can be used for juice, wine, and tabletop grapes. The quality that must be maintained is vastly different for each of these. The ripeness and presentation of tabletop grapes differ from the other two uses; hence, quality control is critical. Deciding which grapes are ready to pick is a labor-intensive approach, and one must ensure the maturity, consistency, and quality of grapes. Making this decision visually requires expert training, which is not easily scalable. Using robotic harvesting instead of human pickers is being explored. Robots can achieve these goals using sensors that will keep the quality in view while speeding the process.

Compiled from L. Rosencrance. (2018, May 31). "Tabletop Grapes to Get Picked by Robots in India, with Help from Virginia Tech." RoboticsBusinessReview. **https://www.roboticsbusinessreview.com/agriculture/tabletop-grapes-picked-robots-india-virginia-tech/** (accessed September 2018); "Tabletop Grapes to Get Picked by Robots in India." **Agtechnews.com**. **http://agtechnews.com/Ag-Robotics-Technology/Tabletop-Grapes-to-Get-Picked-by-Robots-in-India.html** (accessed September 2018).

Robots in the Defense Industry

For obvious reasons, the military has invested in robotic applications for a long time. Robots can replace humans in places where risk of loss of human life is too great. Robots can also reach areas where humans may not be able to go due to extreme conditions – heat, water, and so forth. Besides the recent growth of drones in military applications, several specific robots have been developed over a long time. Some are highlighted in the next sections.

MAARS MAARS (Modular Advanced Armed Robotic System) is an upgraded version of special weapons observation reconnaissance detection system (SWORDS) robots that were used by the U.S. military during the Iraq war. It is designed for reconnaissance, surveillance, and target acquisition and can have a 360-degree view. Depending on the circumstances, MAARS can drape much firepower into its tiny frame. A variety of ammunition such as tear gas, nonlethal lasers, and grenade launcher can be wrapped in it. MAARS is an army robot that can fight autonomously thus reducing risk to soldiers' lives while also protecting itself. This robot has seven types of sensors to track the heat signature of an enemy during the day and night. It uses night vision cameras to monitor enemy activities during the night. On command, MAARS fires at opponents. Its other uses include moving heavy loads from one place to another. It provides a range of options from nonlethal force such as warning of an attack. It can also form a two-sided communication system. The robot can also use less lethal weapons such as laughing gas, pepper spray, and smoke and start clusters to disperse crowds. The robot can be controlled from about one kilometer and is designed to increase or decrease speed, climb stairs, and walk on nonpaved roads using wheels rather than tracks.

Compiled from T. Dupont. (2015, October 15). "The MAARS Military Robot." Prezi. **https://prezi.com/fsrlswo0qklp/the-maars-military-robot/** (accessed September 2018); Modular Advanced Armed Robotic System. (n.d.). Wikipedia. **https://en.wikipedia.org/wiki/Modular_Advanced_Armed_Robotic_System** (accessed September 2018); "Shipboard Autonomous Firefighting Robot – SAFFiR." (2015, February 4). YouTube. **https://www.youtube.com/watch?time_continue=252&v=K4OtS534oYU** (accessed September 2018).

SAFFIR (SHIPBOARD AUTONOMOUS FIREFIGHTING ROBOT) Fire on a ship is one of the greatest risks to shipboard life. Shipboard fires have a different and crucial set of problems. Because of the confined space, there are challenges regarding smoke, gas, and limited ability to escape. Even though procedures like fire drills, onboard alarms, fire extinguishers, and other measures provide ways of dealing with fire on the sea, modern technology is in place to tackle this threat in a better way. A U.S. Navy team at the Office of Naval Research has developed SAFFiR. It is a 5 foot 10 inch tall robot. It is not designed to be completely autonomous. It has a humanoid robotic structure so that it can pass through confined aisles and other nooks and corners of a ship and climb ladders. The robot has been designed to work with the obstacles in the passageways in a ship. SAFFiR can use protective fire gear such as fire-protective coats, suppressants, and sensors that are designed for humans. Lightweight and low-friction linear actuators improve its efficiency and control. It is equipped with several sensors: regular camera, gas, and infrared camera for night vision and in black smoke. Its body is designed not only to be fire resistant but also to throw extinguishing grenades. It can work for around half an hour without needing a charge. SAFFiR can also balance itself on an uneven surface.

Compiled from K. Drummond. (2012, March 8). "Navy's Newest Robot Is a Mechanized Firefighter." **wired.com**. **https://www.wired.com/2012/03/firefight-robot/** (accessed September 2018); P. Shadbolt. (2015, February 15). "U.S. Navy Unveils Robotic Firefighter." CNN. **https://www.cnn.com/2015/02/12/tech/mci-saffir-robot/index.html** (accessed September 2018); T. White. (2015, February 4). "Making Sailors 'SAFFiR' – Navy Unveils Firefighting Robot Prototype at Naval Tech EXPO." America's Navy. **https://www.navy.mil/submit/display.asp?story_id=85459** (accessed September 2018).

Pepper

Pepper is a semihumanoid robot manufactured by SoftBank Robotics that can understand human emotions. A screen is located on its chest. It can identify frowning, tone of voice, smiling, and user actions such as the angle of a person's head and crossed fingers. This way Pepper can determine if a person's mood is good or bad. Pepper can walk autonomously, recognize individuals, and can even lift their mood through its conversation.

Pepper has a height of 120 cms (about 4 feet). It has three directional wheels attached, enabling it to move all around the place. It can tilt its head and move its arms and fingers and is equipped with two high-definition cameras to understand the environment. Because of its anticollision functionalities, Pepper reduces unexpected collisions and can recognize humans as well as obstacles nearby. It can also remember human faces and accepts smartphone and card payments. Pepper supports commands in Japanese, English, and Chinese.

Pepper is deployed in service industries as well as homes. It has several advantages for effectively communicating with customers but has also been criticized at places for incompetence or security issues. The following examples provide information on its applications and drawbacks:

- Interacting with robots while shopping is changing the face of AI in commercial settings. Nestlé Japan, a leading coffee manufacturer, has employed Pepper to sell Nescafé machines to enhance customer experience. Pepper can explain the range of products Nestlé has to offer and recognize human responses using facial recognition and sounds. Using a series of questions and responses to them, the robot identifies a consumer's need and can recommend the appropriate product.
- Some hotels such as Courtyard by Marriott and Mandarin Oriental are employing Pepper to increase customer satisfaction and efficiency. The hotels use Pepper to increase customer engagement, guide guests toward activities that are taking

place, and promote their reward programs. Another goal is to collect customer data and fine-tune the communication according to customer preferences. Pepper was deployed steps away from the entry at Disneyland theme park hotels, and it immediately increased customer interactions. Hotels use Pepper to converse with guests while they are checking in or out or to guide them to the spa, gym, and other amenities. It can also inform guests about campaigns and promotions and help staff members avoid the mundane task of enrolling guests in a loyalty program. Customer reactions are largely quite positive in regard to this.

- Central Electric Cooperative (CEC), an electric distribution cooperative located in Stillwater, Oklahoma, has installed Pepper to monitor outages. CEC serves more than 20,000 customers in seven counties in Oklahoma. Pepper is connected to the operations center to read information about live outages, and by connecting them to geographic information system (GIS) maps it can also inform operations about the live locations of service trucks. At CEC, Pepper is also used for conferences where attendees can know more about the company and its services. Pepper answers a range of questions regarding energy consumption. In the future, the company plans to invest more in robots to meet its requirements. See Figure 10.2 that shows Pepper participating as a team member during a prospective employee interview to provide input about CEC's programs and so on.
- Fabio, a Pepper robot, was installed as a retail assistant at an upmarket food and wine store in England and Scotland. A week after implementing it, the store pulled the service because it was confusing customers, and they preferred the service from personal staff rather than Fabio. It provided generic answers on queries such as the shelf location of items. However, it failed to understand completely what the customer was requesting due to background noise. Fabio was provided another chance by placing it in a specific area that attracted only a few customers. Then they also complained about Fabio's inability to move around the supermarket and direct them to a specific section. Surprisingly, the staff at the market became accustomed to Fabio rather than considering it as a competitor.

FIGURE 10.2 **Pepper Robot as a Participant in a Group Meeting.** *Source:* Central Electric Cooperative.

- Pepper has several security concerns that were pointed out by Scandinavian researchers. According to them, it is easy to have unauthenticated root-level access to the bot. They also found the robot to be prone to brute force attack. Pepper's functions can be programmed using various application programming interfaces (APIs) through languages such as Python, Java, and C++. This feature can cause it to provide access to all its sensors, making it not secure. An attacker can establish a connection and then use Pepper's mic, camera, and other features to spy on people and their conversations. This is an ongoing issue for many robots and smart speakers.

Compiled from "Pepper Humanoid robot helps out at hotels in two of the nation's most-visited destinations (2017)". SoftBank Robotics. **https://usblog.softbankrobotics.com/pepper-heads-to-hospitality-humanoid-robot-helps-out-at-hotels-in-two-of-the-nations-most-visited-destinations** (accessed November 2018); R. Chirgwin. (2018, May 29). "Softbank's 'Pepper' Robot Is a Security Joke." The Register. **https://www.theregister.co.uk/2018/05/29/softbank_pepper_robot_multiple_basic_security_flaws/** (accessed September 2018); A. France. (2014, December 1). "Nestlé Employs Fleet of Robots to Sell Coffee Machines in Japan." *The Guardian*. **https://www.theguardian.com/technology/2014/dec/01/nestle-robots-coffee-machines-japan-george-clooney-pepper-android-softbank** (accessed September 2018); Jiji. (2017, November 21). "SoftBank Upgrades Humanoid Robot Pepper." *The Japan Times*. **https://www.japantimes.co.jp/news/2017/11/21/business/tech/softbank-upgrades-humanoid-robot-pepper/#.W6B3qPZFzIV** (accessed September 2018); C. Prasad. (2018, January 22). "Fabio, the Pepper Robot, Fired for 'Incompetence' at Edinburgh Store." *IBN Times*. **https://www.ibtimes.com/fabio-pepper-robot-fired-incompetence-edinburgh-store-2643653** (accessed September 2018).

Da Vinci Surgical System

Over the last decade, the use of robotics has emerged in surgeries. One of the most famous robotic systems used in surgery is the Da Vinci system that has performed thousands of surgeries. According to surgeons, Da Vinci is the most ubiquitous robot used in more units than any other robot. It is designed to perform numerous nominally invasive operations and can perform simple as well as complex and delicate surgeries. The critical components of Da Vinci are the surgeon console, patient side cart, endowrist instruments, and vision system.

The surgeon console is where the surgeon operates the machine. It provides a high-definition, 3D image of the inside of the patient's body. The console has master controls that a surgeon can grasp by the robotic fingers and operate on the patient. The movements are accurate and in real time, and the surgeon is entirely in control and can prevent the robotic fingers from moving by themselves. The patient side cart is the location where the patient resides during the operation. It has either three or four arms attached that the surgeon controls using master controls, and each arm has certain fixed pivot points around which the arms move. The third component is the endowrist instruments, which are available while performing surgery. They have a total of seven degrees of freedom, and each instrument is designed for a specific purpose. Levers can be released quickly for a change of instruments. The last component is a vision system, which has a high-definition, 3D endoscope and image-processing device that provides real-life images of the patient's anatomy. A viewing monitor also helps the surgeon by providing a broad perspective during the process.

Patients who have surgery that used the Da Vinci system heal faster than those performed by traditional methods because the cuts by robotic arms are quite small and precise. A surgeon must undergo online and hands-on training and must perform at least five surgeries in front of a surgeon who is certified to use the Da Vinci system. This technology does increase the cost of the surgery, but its ability to ease pain while increasing precision makes it the future of such procedures.

Compiled from "Da Vinci Robotic Prostatectomy – A Modern Surgery Choice!" (2018). Robotic Oncology. **https://www.roboticoncology.com/da-vinci-robotic-prostatectomy/** (accessed September 2018); "The da Vinci® Surgical System." (2015, September). Da Vinci Surgery. **http://www.davincisurgery.com/da-vinci-surgery/da-vinci-surgical-system/** (accessed September 2018).

Snoo – A Robotic Crib

Snoo, a robotic, Wi-Fi-enabled crib was developed by Yves Behar, pediatrician Dr. Harvey Karp, and MIT–trained engineers. According to its designers, Snoo mimics Dr. Karp's famous sleep strategy called the five S's, which implies swaddled, side or stomach position, shush, swing, and suck. Snoo is an electrified crib that puts babies to sleep automatically. It recreates sensations experienced by the child during the last trimester of a pregnancy. Infants are at maximum ease when they hear white noise, feel movements, and are wrapped, which Snoo provides at par. Once a baby is securely attached to the bassinet, Snoo senses whether it is fussy, keeps track of its movements, and, if found, moves the crib in a womblike motion until the baby calms down. An app can be installed on Snoo's smartphone to control its speed and white noise. Also, Snoo can be turned off after eight minutes or can continue rocking through the night. The company advertises it as the safest bed ever made with a built-in swaddling strap that ensures that the child does not move from his or her back. Snoo prevents parents from getting up several times in the night to do this themselves; hence, it gives them a sound sleep.

Compiled from S. M. Kelly. (2017, August 10). "A Robotic Crib Rocked My Baby to Sleep for Months." CNN Tech. **https://money.cnn.com/2017/08/10/technology/gadgets/snoo-review/index.html** (accessed September 2018); L. Ro. (2016, October 18). "World's First Smart Crib SNOO Will Help Put Babies to Sleep." Curbed. **https://www.curbed.com/2016/10/18/13322582/snoo-smart-crib-yves-behar-dr-harvey-karp-happiest-baby** (accessed September 2018).

MEDi

MEDi, short for Machine and Engineering Designing Intelligence, is available at six hospitals in Canada and one in the United States. MEDi helps reduce stress in children from painful surgeries, tests, and injections. It is two feet tall and weighs around 11 pounds. It looks like a toy. Dr. Tanya Beran proposed using MEDi after working in hospitals where she heard children exclaiming with joy at the sight of the robot. She suggests that since there is not enough pain management expertise available in such situations, technology can provide a helping hand. The robot can speak 19 languages and can easily be integrated into various cultures. Aldebaran built this robot, which calls itself NAO. It can cost $8,000 and more. Beran bought MEDi to life by adding software that could operate in hospital settings with kids. MEDi strikes up conversations with the kids during a variety of procedures. It was first programmed for flu vaccines and since has been used in other tests. MEDi can even tell story to a child. The robot helps not only children but also nurses by lowering children's stress and relaxing them. Parents have said that when children leave the hospital, they did not speak about needles and pain but in fact left with happy memories.

Compiled from A. Bereznak. (2015, January 7). "This Robot Can Comfort Children Through Chemotherapy." Yahoo Finance. **https://finance.yahoo.com/news/this-robot-can-comfort-children-through-107365533404.html** (accessed September 2018); R. McHugh & J. Rascon. (2015, May 23). "Meet MEDi, the Robot Taking Pain Out of Kids' Hospital Visits." NBC News. **https://www.nbcnews.com/news/us-news/meet-medi-robot-taking-pain-out-kids-hospital-visits-n363191** (accessed September 2018).

Care-E Robot

Airports are growing in size and the number of people who go to them, and this has increased air traffic, flight cancellations, and gate switches, causing travelers to run to different boarding gates. KLM Royal Dutch Airlines is trying a new way to ease this process from problems related to security, boarding gates, and hectic travel with the use of the blue bot "Care-E Robot." This service is scheduled to launch at international airports in New York and

San Francisco. This robot could be found at security checkpoints and take travelers and their carry-on luggage wherever they need to go. Through its nonverbal sounds and signals, Care-E directs travelers to scan their boarding passes and, once scanned, is at their service when they are busy strolling the shops or using restrooms. Care-E also avoids collision using eight sensors with its "peripheral collision avoidance." One of its best features is to relate boarding gate changes to travelers and provide them transportation to the newly assigned gate.

Care-E Robot can carry luggage weighing up to 80 pounds. It runs at a speed of 3 mph, which might be a little too slow for someone running late to catch a flight. However, early travelers who want to explore the airport can use Care-E on a free trial for two days. Implementations of robots like these have not yielded the desired results due to frequent changes in airport policies regarding batteries, but the market for such a robot is quite optimistic about its future.

Compiled from M. Kelly. (2018, July 16). "This Adorable Robot Wants to Make Air Travel Less Stressful." The Verge. **https://www.theverge.com/2018/7/16/17576334/klm-royal-dutch-airlines-robot-travel-airport** (accessed September 2018); S. O'Kane. (2018, May 17). "Raden is the Second Startup to Bite the Dust After Airlines Ban Some Smart Luggage." Circuit Breaker. **https://www.theverge.com/circuitbreaker/2018/5/17/17364922/raden-smart-luggage-airline-ban-bluesmart** (accessed September 2018).

AGROBOT

The combination of sweetness loaded with multiple health benefits makes strawberries one of the world's most popular and consumed fruits. Close to 5 million tons of strawberries are harvested every year, an upward trend in the United States, Turkey, and Spain as top harvesters. AGROBOT, a company engaged in the business of agricultural robots, has developed a robot that can harvest strawberries at any place. Robots using 24 robotic manipulators built on a mobile platform work to identify superior quality strawberries.

Strawberries require a high degree of care because they are delicate compared to other fruits. Fruits such as apples, bananas, and mangoes ripen after being picked whereas strawberries are picked at their full maturity. Hence, harvesting strawberries has been an entirely manual process until recently. AGROBOT was developed in Spain; this robot performs automated processes except selecting the strawberries and packing them. To protect strawberries from being squeezed during picking, the robot cuts them with two razor-sharp blades and catches them in baskets lined with rubber rolls. Once full, the baskets are placed on a conveyor belt and passed to the packing station. Human operators can directly select and pack the berries.

AGROBOT is operated by one man, and a maximum of two people can ride on it. Robotic arms control the coordination between blades and basket. The robot has four main components: inductive sensors, ultrasonic sensors, a collision control system, and a camera system. Camera-based sensors view each fruit and analyze it for ripeness according to its form and color; once a berry is ripe, the robot cuts it from its branches with precise movements. Each arm is fortified with two inductive sensors to stop at the end positions. The collision control system must be capable of responding to dust, temperature change, vibration, and shock; hence, an ultrasonic sensor is attached to the robot to prevent the arms from touching the ground. Each wheel is equipped with ultrasonic sensors to determine the distance between the strawberry and the robot's current position. These sensors also help in keeping the robot on track and preventing damage to the fruit. Signals received from the sensors are continuously transmitted to an automatic steering system to regulate the position of wheels.

Compiled from "Berry Picking at Its Best with Sensor Technology." Pepperl+Fuchs. **https://www.pepperl-fuchs.com/usa/en/27566.htm** (accessed September 2018); R. Bogue. (2016). "Robots Poised to Revolutionise Agriculture." *Industrial Robot: An International Journal, 43*(5), pp. 45–456; "Robots in Agriculture." (2015, July 6). Intorobotics. **https://www.intorobotics.com/35-robots-in-agriculture/** (accessed September 2018).

SECTION 10.4 REVIEW QUESTIONS

1. Identify applications of robots in agriculture.
2. How could a social support robot such as Pepper or MEDi be useful in healthcare?
3. Based on the illustrative applications of robots in this section, build a matrix where the rows are the robots' capabilities and the columns are industries. What similarities and differences do you observe across these robots?

10.5 COMPONENTS OF ROBOTS

Depending on their purpose, robots are made of different components. However, all robots have some common ones, and others are tweaked according to a robot's purpose. Figure 10.3 identifies the components. Common components of the robots are described next.

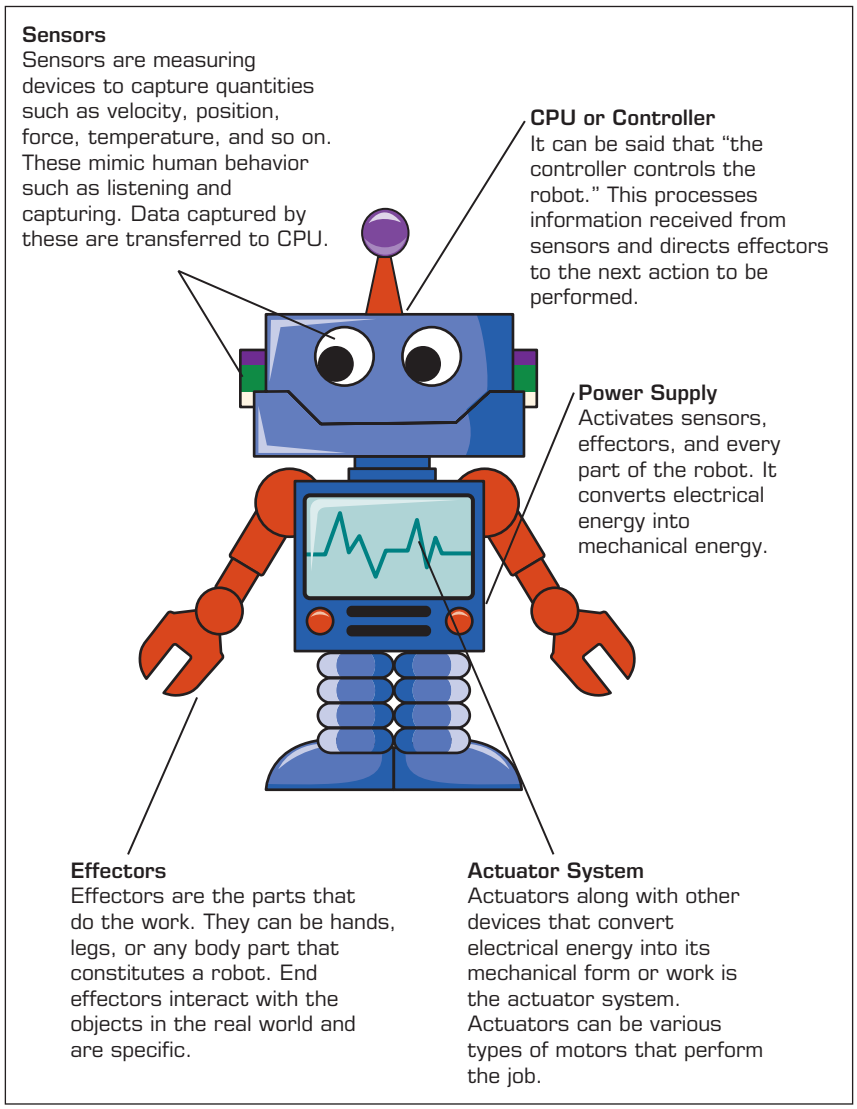


FIGURE 10.3 General Components of a Robot.

POWER CONTROLLER A power controller is the driving force of a robot. Most robots run on batteries, but a few are powered by a direct current (DC) electrical supply. Other factors (i.e., usage, sufficient power to drive all parts) must be kept in mind while designing robots.

SENSORS Sensors are used to direct a robot in its surrounding. Force sensors, ultrasound sensors, distance sensors, laser scanners, and so on help robots to make decisions according to their environment. Sensors are used for robots to identify speech, vision, temperature, position, distance, touch, force, sound, and time. Vision sensors or cameras are used to build a picture of the environment and for the robot to learn about it and to differentiate between which items to choose and which to ignore. In collaborative robots, sensors are also used to prevent them from bumping into humans or other robots. This way, humans and robots can work next to each other without the fear that the robot might unintentionally harm the human. Sensors collect information and send it to the central processing unit (CPU) electronically.

EFFECTORS OR ROVER OR MANIPULATOR An effector is nothing but a body of a robot. It can also describe the devices that affect the environment, such as hands, legs, arms, bodies, and fingers. The CPU controls the actions of effectors. An essential function of them is to move the robot and other objects from one place to another, and their characteristics depend on the role that has been outlined. Industrial robots have end effectors that contribute to the robot's work as a hand. Depending on the type of robot, end effectors can be magnets, welding torches, or vacuums.

NAVIGATION OR ACTUATOR SYSTEM Actuators are devices that define how a robot travels. With the help of an actuator, electrical energy converts into mechanical energy, enabling the robot to move back, forward, left, right and to lift, drop, and perform its job. The actuator can be a hydraulic cylinder or an electric motor. The actuator system is the way that all of the robot's components are embedded into one.

CONTROLLER/CPU This is the brain of the robot and has the AI embedded in it. The CPU allows a robot to perform its function by connecting all systems into one. It also provides commands for the robot to learn from the surrounding movement of the body or any of its actions.

SECTION 10.5 REVIEW QUESTIONS

1. What are the common components of a robot?
2. What is the function of sensors in a robot?
3. How many different types of sensors might exist in a robot?
4. What is the function of a manipulator?

10.6 VARIOUS CATEGORIES OF ROBOTS

Robots perform a variety of functions. Depending on these, robots can be categorized into the following categories.

PRESET ROBOTS Preset robots are preprogrammed. They have been designed to perform the same task over time and can work 24 hours a day, 7 days a week without any breaks. Preset robots do not alter their behavior. Therefore, these robots have an incredibly low error rate and are suitable for wearisome work. They are frequently used in manufacturing sectors such as the mobile industry, vehicle manufacturing, material handling, and welding to save time and money. Preset robots deliver jobs in environments where it is hazardous

for humans to work. Robots move heavy objects, perform assembly tasks, paint, inspect parts, and handle chemicals. A preset robot articulates according to the operation it performs. It can perform a significant role in the medical field because the tasks it performs must have high efficiency at a level comparable to human beings.

COLLABORATIVE ROBOTS OR COBOTS Cobots are the robots that can collaborate with human workers, assisting them to achieve their goals. The use of cobots is trending in the market, and there is an excellent outlook for collaborative robots. According to the survey by MarketsandMarkets, the cobots market in 2020 will be worth around $3.3 billion. There are various functions of collaborative robots. Depending on the usage, the collaborative robots are used. Collaborative robots have various applications in manufacturing as well as the medical industry.

STAND-ALONE ROBOTS Stand-alone robots are the robots that have a built-in AI system and work independently without much interference from humans. These robots perform tasks depending on the environment and adapt to changes in it. With the use of AI, a stand-alone robot learns to modify its behavior and excel in performing its assignment. Autonomous robots have household, military, education, and healthcare applications. They can walk like a human being, avoid obstacles, and provide social-emotional support. Some of these robots are used for domestic purposes as stand-alone vacuum cleaners, such as iRobot Roomba. Stand-alone robots are also used in hospitals to deliver medications, keep track of patients who are yet to receive them, and send this information to the nurses working on that shift and other shifts without chance of any error.

REMOTE-CONTROLLED ROBOTS Even though robots can perform stand-alone tasks, they do not have human brains; hence, many tasks require human supervision. These robots can be controlled via Wi-Fi, Internet, or satellite. Humans direct remote-controlled robots to perform complicated or dangerous tasks. The military uses these robots to detonate bombs or to act as soldiers around the clock on the battlefield. In the space program research field, their scope of use is extensive. Remote-controlled cobots are also used to perform marginally invasive surgeries.

SUPPLEMENTARY ROBOTS Supplementary robots enhance the existing capabilities or replace capabilities that a human has lost or does not have. This type of robot can be directly attached to a human's body. It connects to a user's body and communicates with the robot's operator directly or when the operator grips the body. The robot can be controlled by a human body, and in some cases, even by thinking of a specific action. Its applications include serving as a robotic prosthetic arm or providing precision for the surgeons. Extensive research on building prosthetic limbs is being conducted.

SECTION 10.6 REVIEW QUESTIONS

1. Identify some key categories of robots.
2. Define and illustrate the capabilities of a cobot.
3. Distinguish between a preset robot and a stand-alone robot. Give examples of each.

10.7 AUTONOMOUS CARS: ROBOTS IN MOTION

A robot that may eventually touch most people's lives is an autonomous (self-driving) car. Like many other technologies, self-driving cars have been at peak hype recently, but people also recognize their technical, behavioral, and regulatory challenges. Nevertheless, technology and processes are evolving to make the self-driving car a reality in the future,

at least in specific settings if not all over the world. Early versions of self-driving cars were enabled by the radio antenna developed in 1925. In 1989, researchers at Carnegie Mellon used neural networks to control an autonomous vehicle. Since then, many technologies have come together to accelerate development of self-driving cars. These include:

- **Mobile phones:** With the help of low-powered computer processors and other accessories such as cameras, mobile phones have become ubiquitous. Many technologies developed for phones, such as location awareness and computer vision, are finding applications in cars.
- **Wireless Internet:** Connectivity has become much more feasible with the rise of 4G networks and Wi-Fi. Going forward, growth in 5G will perhaps be important for self-driving cars to allow their processors to communicate with each other in real time.
- **Computer centers in cars:** A number of new technologies are available in today's cars, such as rearview cameras and front and back sensors that help vehicles detect objects in the environment and alert the driver to them or even take necessary actions automatically. For example, adaptive cruise control automatically adjusts the speed of a car based upon the speed of the vehicle in front.
- **Maps:** Navigation maps on mobile phones or navigation systems in cars have made a driver's job easy with regard to navigation. These maps enable an autonomous vehicle to follow a specific path.
- **Deep learning:** With advances in deep learning, the ability to recognize an object is a key enabler of self-driving cars. For example, being able to distinguish a person from an object such as a tree, or whether the object is moving or stationary is critical in taking actions in a moving vehicle.

Autonomous Vehicle Development

The heart of an autonomous vehicle system is a laser rangefinder (or light detection and ranging – lidar device), which is on the vehicle's roof. The lidar generates a 3D image of the car's surroundings and then combines it with high-resolution world maps to produce different data models for taking action to avoid obstacles and follow traffic rules. In addition, many other cameras are mounted. For example, a camera positioned near a rearview mirror detects traffic lights and takes videos. Before making any navigation decisions, the vehicle filters all data collected from the sensor and camera and builds a map of its surroundings and then precisely locates itself in that map using GPS. This process is called *mapping and localization*.

The vehicle also consists of other sensors such as the four radar devices that are on the front and back bumpers. These devices allow the vehicle to see far distances so that they can make decisions beforehand and deal with fast-moving traffic. A wheel encoder determines the vehicle's location and maintains records of its movements. Algorithms such as neural networks, rule-based decision making, and a hybrid approach are used to determine the vehicle's speed, direction, and position, and the collected data are used to direct the vehicle on the road to avoid obstacles.

Autonomous vehicles must rely on detailed maps of roads. Thus, before sending driverless cars on roads, engineers drive a route several times and collect data about its surroundings. When driverless vehicles are in operation, engineers compare the data acquired by them to the historical data.

There is an entire town built for the sole purpose of testing autonomous vehicles. It is located in Michigan. This city has no single resident, and self-driving vehicles roam the streets without the risks in the real world. This city, called Mcity, is truly a city for robotic vehicles. Mcity includes intersections, traffic signals, buildings, construction work, and

moving obstacles such as humans and bicycles similar to those in real cities. Autonomous vehicles are not only tested in this closed environment but are being used in the real world as well.

Google's Waymo unit is one of the early pioneers of self-driving vehicles. They have been tested on California roads, but before they start to drive next to human-driven cars, companies have to test them thoroughly because one negative incident can impede their acceptance. For example, in the spring of 2018, a self-driving vehicle being tested by Uber killed a pedestrian in Tempe, Arizona. This led to the suspension of all public testing of autonomous vehicles by Uber. The technology is still in development, but it has come far enough that limited testing on public roads is safe. We might be surprised in the near future by the fact that the person in the driver's seat of a vehicle next to you in traffic might not actually be driving it at all.

In 2016, the U.S. Department of Transportation (DOT) began to embrace driverless vehicles to speed their development. In September 2016, DOT announced the first-ever guidelines for autonomous driving. A groundbreaking announcement by the National Highway Traffic Safety Administration (NHTSA) a month later allowed for the AI system controlling Google's self-driving vehicle to be considered a driver in response to the company's proposal to the NHTSA in November 2015.

Some states currently have specific laws that ban autonomous driving. For example, as of this writing, the state of New York does not allow any hands-free driving. Without clear regulations, testing self-driving vehicles is a challenge. Although a few states such as Arizona, California, Nevada, Florida, and Michigan currently allow autonomous vehicles on the road, California is the only one with licensing regulations at this point.

Google might be the most well known for autonomous vehicles, but it is not the only one. A handful of the most powerful companies, such as Uber and Tesla, are in the same race as well. Every major car company is working either with technology companies or its own technology to develop autonomous vehicles or at least to participate in this revolution.

Issues with Self-Driving Cars

Autonomous cars have been connected to a number of issues.

- **Challenges with technology:** There have been several challenges with the technology used in self-driven cars. Several software and mechanical hurdles are still to be overcome in order to roll out a fully autonomous car. For example, Google is still trying to update its software on an almost daily basis for its self-driven car. Several other companies are still trying to figure out the amount of authority to be transferred when a human driver takes control from an automatic vehicle.
- **Environmental challenges:** Technology and mechanical capabilities cannot yet address many environmental factors affected by self-driving cars. For example, there are still concerns regarding their performance in bad weather. Likewise, several systems have not been tested in extreme conditions such as snow and hail. There are several tricky navigating situations on the road, such as when an animal jumps onto it.
- **Regulatory challenges:** All companies planning to become involved with self-driving cars need to address regulatory hurdles. There are still many unanswered questions about the regulation of autonomous driving. Several questions about liability include these: What will a license involve? Will new drivers be required to get traditional licenses even if they are not drivers? What about young people, or older people with disabilities? What will be required to operate these new vehicles? Governments need to work quickly to catch up with the booming technology.

Considering that public safety is on the line, auto regulations should be some of the strictest regulations in the modern world.

- **Public trust issues:** Most people do not yet believe that an autonomous car can keep them safe. Trust and consumer acceptance are the crucial factors. For example, if there is a situation when an autonomous car is being forced to choose between the life of a passenger versus that of a pedestrian, what should be done? Consumers may refuse the whole idea of driverless cars. No technology can be perfect, but the question is which company will be able to best convince its customers to entrust their lives to them.

Advances similar to those for self-driving cars are being explored in other autonomous vehicles. For example, several companies have already launched trials of self-driving trucks. Autonomous trucks, if ever fully deployed, will have a massive disruptive effect on jobs in the transportation industry. Similarly, self-driving tractors are being tested. Finally, autonomous drones and aircrafts are also being developed. These developments will have a huge impact on future jobs while creating other new jobs in the process.

Self-driving vehicles have become part of this world of technology in spite of related technical and regulatory barriers. Autonomous vehicles are yet to achieve the knowledge capabilities of human drivers, but as the technology improves, more-reliable driving vehicles will become a reality. Like many technologies, the short-term impact may be cloudy, but the long-term impact is yet to be determined.

SECTION 10.7 REVIEW QUESTIONS

1. What are some of the key technology advancements that have enabled the growth of self-driving cars?
2. Give examples of regulatory issues in self-driving cars.
3. Conduct online research to identify the latest developments in autonomous car deployment. Give examples of positive and negative developments.
4. Which type of self-driving vehicles are likely to have the most disruptive effect on jobs, and why?

10.8 IMPACT OF ROBOTS ON CURRENT AND FUTURE JOBS

Robotics has been a boon to the manufacturing industry. Besides automation that is possible with robotics, new technologies such as image recognition systems are automating jobs that used to require humans for inspection and quality control.

Various industry experts report that by 2025, up to 25 percent of current jobs will be replaced by robots or AI. Davenport and Kirby's book *Humans Need Apply: Winners and Losers in the Age of Smart Machines* (2016) focuses on this topic. Of course, many other researchers, journalists, consultants, and futurists have given their own predictions. In this section, we review some related issues. These issues are relevant to AI in general and robotics in particular. Thus, Chapter 14 will also cover these issues, but we want to study these in the context of robotics in this chapter.

As a group activity, watch the following video: **https://www.youtube.com/watch?v=GHc63Xgc0-8**. Also watch **https://www.youtube.com/watch?v=ggN8wCWSIx4** for a different view. What are your takeaways from these videos? What is the most likely scenario in your view? How can you prepare for the day when indeed humans may not need to apply for many jobs?

IBM Watson's ability to digest vast amounts of data in the medical research literature and provide the latest information to a physician has been written about in the literature. Similar job enhancement opportunities in many other areas have been seen. Consider

this: AI-powered technologies such as narrative science and automated insights that can ingest structured data include visualizations generated by software such as Tableau and develop an initial draft of a story to narrate what the results convey. Of course, that would appear to threaten the job of a journalist or even a data scientist. In reality, this can also enhance that job by presenting an initial draft of a story. Then the storyteller can focus on more advanced and strategic issues related to that data and visualization.

The power of consistency and comprehensiveness can also be helpful in the completion of jobs. For example, as noted by Meister (2017), chatbots can likely provide much of the initial human resource (HR) information to new employees. Chatbots can also be helpful in providing such information to remote employees. A chatbot is more likely than a human to provide complete and consistent information each time. Of course, this implies that workers whose main job is to recite such information to each new employee or serve as the first source of information may not be needed.

Hernandez (2018) identified seven job categories into which robotics in particular and AI in general will expand. She also quoted several other studies. According to a McKinsey & Co study, AI could result in 20–50 million new jobs in the next 10–15 years. McKinsey also predicts that 75–375 million people may need to change jobs/occupations in the same time period because of robotics and AI. According to Hernandez, the following seven jobs are likely to increase:

1. **AI development:** This is an obvious growing area. As more companies develop products and services based on AI, the need for such developers will continue to increase. As an example, iRobot Co, which produces robotic vacuum cleaners, is shifting its hiring from hardware to software engineers as it works to develop its next generation of products that are more adaptive and AI based. Newer robot vacuums are going to be able to "see" a wall. They can also alert the owner to how long the cleanup took and the area swept.
2. **Customer–robot interactions:** As more companies deploy robots in these organizations, acceptance of such robots by both employees and customers is uncertain. A new job category has emerged to study the interactions between a robot and its coworkers and customers and to retrain the robot or take this information into account in designing the next generation. Clearly, the study of such interactions may enable the use of analytics/data science as well.
3. **Robot managers:** Although robots might do the bulk of their work in a specific situation, humans will still need to observe them and ensure that the work is progressing as expected. Further, if any unusual conditions arise, a human worker has to be alerted and respond to the situation. This would be true in many settings where the robots are performing the bulk of tasks in areas such as manufacturing. Hernandez (2018) gives an example of Cobalt robots, which work as security guards. These robots alert a human whenever an intruder is detected or they notice anything unusual. Of course, a human robot manager is typically able to supervise many more robots than human workers because the primary role of the manager is to supervise them and respond to unusual situations.
4. **Data labelers:** Robots or AI algorithms learn from examples. And the more examples they are given, the better their learning can be (see Chapter 5 for a longer description of this issue). For example, image recognition systems in virtually every setting (see Chapter 5 on deep learning for examples) require as many examples as possible to improve those systems' recognition capability. This is crucial for not just facial recognition but also image applications to detect cancer from X-ray images, weather features from radar images, and so on. It requires that humans view the example images and label them as representing a specific person, feature, or class. This work is tedious and requires humans. Many companies have hired hundreds

of human labelers to view the images and tag them appropriately. As such image applications grow, the need for labelers will also increase. These workers are also needed for continuous improvement of the robot or AI algorithms by recording false positives or newer examples.

5. **Robot pilots and artists:** Robots in general and drones in particular are being used to provide action shots using overhead cameras or angles that would be difficult if not impossible for humans to do. Drones could also be dressed as birds or flowers and provide a unique overview as well as enhance a setting. Similarly, other robots might be dressed in unique outfits to create a cultural ambiance. Such designer/makeup artists are being hired by many companies that provide services for events such as concerts, weddings, and so forth. In addition, drone piloting has become a highly specialized skill for entertainment, commercial, and military applications. These jobs will increase as the applications evolve.
6. **Test drivers and quality inspectors.** Autonomous cars are already becoming reality. With each such automation of vehicles, at least for the foreseeable future, there is a growing need for safety drivers who monitor each vehicle's performance and take appropriate actions in unusual situations. Their jobs would not entail the use of remote controls as drone pilots employ but continuous watch of the vehicle's operations and response to emergent situations. Similar jobs also exist in other robotic applications as the robots are trained and tested to work in specific settings.
7. **AI lab scientists.** This brings us to the very first category of new jobs we identified—AI coders. While their job is to develop the algorithms for robots or AI programs, a similar category of highly specialized users is also emerging—folks who are trained and employed in using these hardware and software systems for special applications. For example, physicians have to undergo additional training to be certified in the use of robots in their surgeries, cardiology and urology practices, and so on. Another category of such specialists involves scientists who customize these robots and AI algorithms for their domain. For example, quite a few companies are using AI tools to identify new drug molecules to develop and test new treatment options for diseases. AI could speed such development. These scientists not only develop their domain expertise but also data scientists' knowledge or at least the ability to work with data scientists to create their new applications.

Although the preceding list identifies several categories of jobs that are likely to develop or increase, millions of jobs are likely to be eliminated. For example, automation is already impacting the number of jobs in logistics. When autonomous trucks become a reality, at least some of the well-paying jobs in transportation will likely be gone. There might be disagreements on when the massive change may occur, but the long-term impact on jobs is certain to occur. The major issue this time is that many of the knowledge economy "white-collar" jobs are the ones that are more likely to be automated. And this change is unprecedented in history. Many social scientists, economists, and leading thinkers are worried about the upheaval that this next wave of robotic automation will cause, and they are considering various solutions. For example, the concept of universal basic income (UBI) has been proposed. UBI proponents argue that giving every citizen a minimal basic income will ensure that no one goes hungry despite the massive loss of jobs that is likely to occur. Others, for example, Lee (2018), have argued that providing UBI may not satisfy human beings' need for meaningful achievements and contributions in life. Lee proposes a social investment stipend (SIS), which would recognize individuals' contributions to society for providing support and care, community service, or education. The stipend would be paid in recognition of an individual's service in one of these categories. Lee's book focuses on this issue and is one of the many ideas being proposed on how to plan for and address the upcoming disruption from automation. Our goal in this section is to simply alert you to these issues.

SECTION 10.8 REVIEW QUESTIONS

1. Which jobs are most at risk of disappearing as the result of the new robotics revolution?
2. Identify at least three new categories of jobs that are likely to result in a significant number of new employees.
3. Are the tasks undertaken by data labelers just for one time or longer lasting?
4. Research the concepts of UBI and SIS.

10.9 LEGAL IMPLICATIONS OF ROBOTS AND ARTIFICIAL INTELLIGENCE

As we noted in the previous section, the impact of AI in general and robotics in particular is far and wide and can be studied both specifically in the context of robots and more broadly for AI. Legal implications of robotics and AI are discussed in this chapter and in Chapter 14. Many legal issues are yet to be untangled as we embrace and employ AI technologies, robots, and self-driving cars. This section highlights some of the key dimensions of legal impacts related to AI. The following material has been contributed by Professor Michael Schuster, assistant professor of Legal Studies at the Spears School of Business at Oklahoma State University. He is a noted expert in legal matters related to AI. He has also published extensively in this area.

Tort Liability

Self-driving cars and other systems controlled by AI represent a Pandora's box of potential tort liability (where a wrongful act creates on obligation to pay damages to another). Imagine that a motorcyclist is injured when a self-driving car veers into his lane and they collide. This was the alleged event that led to *Nilsson v. General Motors* (N.D. California, 2018)—a case with the potential to address difficult questions of AI-created tort liability. The suit settled, and thus did not clarify who should pay when someone is injured by an AI-controlled system. Potential candidates for liability include programmers of an AI system, manufacturers of a product incorporating AI, and owners of a product at the time it harmed another. Medical malpractice litigation may similarly be altered by new technologies. As doctors defer some decision making to AI, lawsuits for injurious medical care move from professional liability cases (against the doctor) to product liability (against manufacturers of AI systems). An early example of this phenomenon is the lawsuits over allegedly botched surgeries using the Da Vinci Surgical Systems robot.

Patents

The introduction of AI systems capable of independent or human-assisted invention raises a variety of questions about patenting these creations. Patents have traditionally been granted for novel inventions that would not have been an obvious improvement of a known technology as viewed through the eyes of an average party working in the relevant field. Accordingly, this standard has traditionally asked whether a new technology would have been obvious to a human, but as AI becomes ubiquitous, the scope of what is obvious expands. If the average person in an industry has access to an AI system capable of inventing new things, many improvements on known technologies can become obvious. Since these improvements would then be obvious, they would no longer be patentable. Inventing AI will thus make it harder for a human to get a patent as such technology becomes more commonplace. Moving beyond human inventions, a host of issues arise regarding AI that can *independently* invent. If a person does not contribute to the invention (but rather merely identifies a goal to be achieved or provides background

data), he or she does not satisfy the statutory threshold to be an inventor (and thus, does not qualify for patent ownership). See Schuster (2018). If AI creates an invention without a human inventor, who owns the patent or should a patent be granted at all? Some assert the U.S. Constitution's mandate that patents can be granted only to "inventors" necessarily requires a human actor, and thus, Congress cannot constitutionally allow the patenting of AI-created technologies. Additional commentators present a variety of policy positions arguing why parties such as the computer's owner, the AI's creator, or others should own patents for computer-created inventions. These issues are yet to be resolved.

Property

A basic tenet of U.S. law is a strong protection of property rights. These values extend to corporate entities that can own both real estate and movable property to the exclusion of all others while shareholders retain some allotted portion of the corporation itself. Analysts are presently addressing to what extent property rights should extend to autonomous AI. If a robot were to engage in work for hire, might it be able to make purchases of goods or realty to further its interests? Could an eccentric octogenarian millionaire leave his entire fortune to a loyal robot housekeeper? Topics of this nature will raise a variety of new legal queries, including intestate passage of AI-owned property at its "death" and the standard for death in a nonbiological entity.

Taxation

Robotics and AI will replace a significant number of jobs presently undertaken by humans. Commentators are divided on whether new technologies will create a number of new jobs equal to those replaced by automation. Should the scope of new jobs fall shy of those lost, it is feasible that tax-based issues will be created. A particular concern deals with federal payroll taxes whereby workers and employers pay taxes premised on wages made by the employee. If the aggregate number of workers and net pay are reduced by job automation, the payroll tax base will be reduced. Given that these taxes are important to the sustainability of various government-run safety net programs (e.g., Social Security), payroll tax shortfalls could have significant societal ramifications. In 2017, Bill Gates (Microsoft cofounder) set forth a proposal to tax robots that are used to automate existing human jobs. This new tax would theoretically supplement extant payroll taxes to ensure continued funding for government programs. Commentators are split on the advisability (or need) for such a tax. A common criticism is that taxing robots discourages technological advancement, which is contrary to the accepted policy of encouraging such endeavors. A satisfactory resolution to this debate is yet to be reached.

Practice of Law

Beyond what the law is or should be, AI will have substantial effects within discrete segments of the practice of law. A prime example of this influence is in the area of document review—the part of a lawsuit where litigants evaluate documents provided by their opponents for relevance to the case. Costs associated with this process can be substantial given that some cases entail review of millions of pages by attorneys who bill hundreds of dollars per hour. Corporate clients looking to reduce costs—and law firms seeking a competitive advantage—have adopted (or intend to adopt) AI-based document review systems to minimize the number of billed hours. Similarly, some firms have adopted industry-specific technologies to create competitive advantage. At least one major law firm instituted the use of an AI-driven system to analyze strengths and weaknesses of its clients' patent portfolios.

Constitutional Law

As the state of AI advances, it continues to move toward "human-level" intelligence. But as it becomes more "personlike," questions arise regarding whether AI should be afforded rights commonly granted to humans. The U.S. Constitution's First Amendment provides for freedoms of speech, assembly, and religion, but should rights of this nature extend to AI? For instance, one might argue that these rights preclude the government from dictating what a robot can say (violating its right to free speech). At first blush, this proposition seems far-fetched, but perhaps it is not. On the issue, it is notable that the Supreme Court of the United States recently extended some free speech rights and religious liberties to corporate entities. Accordingly, there is some domestic precedent for affording constitutional protections to nonhuman actors. Further, in 2017, Saudi Arabia granted citizenship to "Sophia," a humanoid robot created by Hong Kong's Hanson Robotics in 2015. How this issue will resolve itself (domestically and globally) remains to be seen.

Professional Certification

There are many activities for which humans must receive certification issued by a government or professional organization prior to undertaking that act (e.g., the practice of law or medicine). As the state of AI progresses, AI will increasingly be capable of performing these state-regulated endeavors independent of human engagement. With this in mind, standards must be developed to determine whether an AI technology is capable of providing satisfactory service in regulated professional fields. If an autonomous robot is capable of passing a state's bar exam, should it be able to give legal advice without human supervision? To the extent that many professional groups require annual training to maintain competence, how will these policies apply to AI technologies? Is there value in requiring that a computer undertaking legal functions "attend" continuing legal education classes? These issues will be settled as AI begins to carry out work currently done exclusively by human professionals like doctors and lawyers.

Law Enforcement

In addition to policy choices detailing what the law is or should be, AI may influence enforcement of the law. Rapid growth in technology will soon afford police forces access to large amounts of near real-time data and computing capacity to determine where crimes are being committed. Recognized infractions may run the gamut from common public transgressions (e.g., running a red light) to more private acts, such as underreporting income on a tax return. The capacity to recognize such criminal acts on a large scale raises a variety of enforcement questions. Discretion in prosecuting infractions has long been a part of law enforcement. Should this power of choice regarding issues associated with stereotype-based prosecution decisions be delegated to AI systems? Moreover, machine-based enforcement programs have consistently been met with questions of their constitutionality (e.g., using cameras to identify drivers not stopping for red lights). While these arguments have thus far proven unsuccessful, they will likely be relitigated as the practice expands. Beyond enforcement questions, some have raised the possibility of implementing AI in the judiciary. For instance, it has been proposed that data-based sentencing may more successfully achieve targeted goals (e.g., successful education while incarcerated or avoidance of recidivism) than arguably idiosyncratic members of the judiciary. Such a mechanism will, of course, raise potential transparency issues and arguments relating to granting too much power to AI systems.

Regardless of the issues that the last two sections have raised, robotics technologies and applications are evolving rapidly. As managers, you have to continue to think about how to manage these technologies while being fully aware of immediate behavioral and legal issues in implementing the technologies.

SECTION 10.9 REVIEW QUESTIONS

1. Identify some of the key legal issues for robotics and AI.
2. Liability for harm (tort liability) is an obvious early question for any technology. What are some of the key challenges in identifying such liability?
3. Recent news about illegal intervention in elections has led to the discussion about who is responsible for damage control. When chatbots and automated social media systems have the ability to propagate "fake news," who should be required to monitor them and prevent such action?
4. What are some of the law enforcement issues in employing AI?

Chapter Highlights

- Industrial automation brought the first wave of robots, but now the robots are becoming autonomous and finding applications in many areas.
- Robotic applications span industries such as agriculture, healthcare, and customer service.
- Social robots are emerging as well to provide care and emotional support to children, patients, and older adults.
- All robots include some common components: power unit, sensors, manipulator/effector, logic unit/CPU, and location sensor/GPS.
- Collaborative robots are evolving quickly, leading to a category called *cobots*.
- Autonomous cars are probably the first category of robots to touch most consumers.
- Self-driving cars are challenging the limits of AI innovation and legal doctrines
- Millions of jobs are at risk of being lost due to the use of robots and AI, but some new job categories will emerge.
- Robots and AI are also creating new challenges in many legal dimensions.

Key Terms

automation
autonomous car
autonomy
effector
patent
robot
sensor
social robot
tort liability
universal basic income (UBI)

Questions for Discussion

1. Based upon the current state of the art of robotics applications, which industries are most likely to embrace robotics? Why?
2. Watch the following two videos: **https://www.youtube.com/watch?v=GHc63Xgc0-8** and **https://www.youtube.com/watch?v=ggN8wCWSIx4** for a different view on impact of AI on future jobs. What are your takeaways from these videos? What is the more likely scenario in your view? How can you prepare for the day when humans indeed may not need to apply for many jobs?
3. There have been many books and opinion pieces written about the impact of AI on jobs and ideas for societal responses to address the issues. Two ideas were mentioned in the chapter – UBI and SIS. What are the pros and cons of these ideas? How would these be implemented?
4. There has been much focus on job protection through tariffs and trade negotiation recently. Discuss how and why this focus may or may not address the job changes coming due to robotics and AI technologies.
5. Laws rely on incentive structures to encourage prosocial behavior. For example, criminal law encourages compliance by punishing those who break the law. Patent law incentivizes creation of new technologies by offering inventors a period of limited monopoly during which they can exclusively use their invention. To what extent do these (and other) incentives make sense when applied to AI? How can incentive structures be

created to encourage AI devices to behave in prosocial manners?

6. To what extent do extralegal considerations come into play with regard to the above issues? Are there moral (or religious) dimensions to be considered when determining whether AI should be given rights similar to those of a person? Would AI-assisted law enforcement or court action erode faith in the criminal justice system and judiciary?
7. Adopting policies that maximize the value of AI encourages future development of these technologies. Such a course, however, is not without drawbacks. For instance, determining that a "robot tax" is not a preferred policy choice would increase the incentive to adopt a robot workforce and improve any relevant technologies. Elevating the state of robotics is a laudable goal, but in this instance, it would come at the anticipated cost of reduced public funds. How should trade-offs such as these be evaluated? Where should encouragement of technological progress (especially regarding AI) fall in the hierarchy of government priorities?

Exercises

1. Identify applications other than those discussed in this chapter where Pepper is being used for commercial and personal purposes.
2. Go through specifications of MAARS at **https://www.qinetiq-na.com/wp-content/uploads/brochure_maars.pdf**. What are the functions of MAARS?
3. Conduct online research to find at least one new robotics application in agriculture. Prepare a brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.
4. Conduct online research to find at least one new robotics application in healthcare. Prepare a brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.
5. Conduct online research to find at least one new robotics application in customer service. Prepare a brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.
6. Conduct online research to find at least one new robotics application in an industry of your choice. Prepare a brief summary of your research: the problem addressed, technology summary, results achieved if any, and lessons learned.
7. Conduct research to identify the most recent developments in self-driving cars.
8. Conduct research to learn and summarize any new investments and partnerships in self-driving cars.
9. Conduct research to identify any recent examples of legal issues regarding self-driving cars.
10. Conduct research to identify any other new types of jobs that would be enabled by AI and robotics beyond what was covered in the chapter.
11. Conduct research to report on the latest projections for job losses due to robotics and AI.
12. Identify case stories for each of the legal dimensions identified by Schuster (2018) in Section 10.9.

References

"A Brief History of Robotics since 1950." **Encyclopedia.com**. **http://www.encyclopedia.com/science/encyclopedias-almanacs-transcripts-and-maps/brief-history-robotics-1950** (accessed September 2018).

Ackerman, E. (2016). *IEEE Spectrum*. **http://spectrum.ieee.org/automaton/robotics/home-robots/tega-mit-latest-friendly-squishable-social-robot** (March 5, 2017).

"Adidas's High-Tech Factory Brings Production Back to Germany." (2017, January 14). *The Economist*. **https://www.economist.com/business/2017/01/14/adidass-high-tech-factory-brings-production-back-to-germany** (accessed September 2018).

Allinson, M. (2017, March 4). "BMW Shows Off Its Smart Factory Technologies at Its Plants Worldwide." Robotics and Automation. **https://roboticsandautomationnews.com/2017/03/04/bmw-shows-off-its-smart-factory-technologies-at-its-plants-worldwide/11696/** (accessed September 2018).

Aoki, S., et al. (1999). "Automatic Construction Method of Tree-Structural Image Conversion Method ACTIT." *Journal of the Institute of Image Information and Television Engine, 53*(6), pp. 888–894 (in Japanese).

"A Robot Cooks Burgers at Startup Restaurant Creator." (2018). Techcrunch. **https://techcrunch.com/video/a-robot-cooks-burgers-at-startup-restaurant-creator/** (accessed September 2018).

Ayres, R., & S. Miller. (1981, November). "The Impacts of Industrial Robots." Report CMU-RI-TR-81-7. Pittsburgh, PA: The Robotics Institute at Carnegie Mellon University.

Bereznak, A. (2015, January 7). "This Robot Can Comfort Children Through Chemotherapy." Yahoo Finance. **https://finance.yahoo.com/news/this-robot-can-comfort-children-through-107365533404.html** (accessed September 2018).

"Berry Picking at Its Best with Sensor Technology." (2018). Pepperl+Fuchs. **https://www.pepperl-fuchs.com/usa/en/27566.htm** (accessed September 2018).

Bogue, R. (2016). "Robots Poised to Revolutionise Agriculture." *Industrial Robot: An International Journal, 43*(5), pp. 450–456

Broekens, J., M. Heerink, & H. Rosendal. (2009). "Assistive Social Robots in Elderly Care: A Review." *Gerontechnology, 8*, pp. 94–103 doi: 10.4017/gt.2009.08.02.002.00.

Carlsson, B. (1998) "The Evolution of Manufacturing Technology and Its impact on Industrial Structure: An International Study." IUI Working Paper 203. Internation Joseph A. Schumpeter Society Conference on Evolution of Technology and Market in an International Context. The Research Institute of Industrial Economics (IUI), Stockholm, May 24–28, 1988.

"Case Study Pepper, Courtyard Marriott." SoftBank Robotics. **https://www.softbankrobotics.com/us/solutions/pepper-marriott** (accessed September 2018).

Chirgwin, R. (2018, May 29). "Softbank's 'Pepper' Robot Is a Security Joke." The Register. **https://www.theregister.co.uk/2018/05/29/softbank_pepper_robot_multiple_basic_security_flaws/** (accessed September 2018).

Coxworth, B. (2018, May 29). "Restaurant Keeps Its Prices Down – With a Robotic Kitchen." New Atlas. **https://newatlas.com/spyce-restaurant-robotic-kitchen/54818/** (accessed September 2018).

"Da Vinci Robotic Prostatectomy – A Modern Surgery Choice!" Robotic Oncology. **https://www.roboticoncology.com/da-vinci-robotic-prostatectomy/** (accessed September 2018).

Drummond, K. (2012, March 8). "Navy's Newest Robot Is a Mechanized Firefighter." **wired.com**. **https://www.wired.com/2012/03/firefight-robot/** (accessed September 2018).

Dupont, T. (2015, October 15). "The MAARS Military Robot." Prezi. **https://prezi.com/fsrlswo0qklp/the-maars-military-robot/** (accessed September 2018).

Engel, J. (2018, May 3). "Spyce, MIT-Born Robotic Kitchen Startup, Launches Restaurant: Video." Xconomy. **https://www.xconomy.com/boston/2018/05/03/spyce-mit-born-robotic-kitchen-startup-launches-restaurant-video/** (accessed September 2018).

Fallon, S. (2015). "A Blue Robotic Bear to Make Sick Kids Feel Less Blue." **https://www.wired.com/2015/03/blue-robotic-bear-make-sick-kids-feel-less-blue/** (accessed August 2018).

Forrest, C. (2015). "Chinese Factory Replaces 90% of Humans with Robots, Production Soars." TechRepublic. **https://www.techrepublic.com/article/chinese-factory-replaces-90-of-humans-with-robots-production-soars/** (accessed September 2018).

France, A. (2014, December 1). "Nestlé Employs Fleet of Robots to Sell Coffee Machines in Japan." *The Guardian*. **https://www.theguardian.com/technology/2014/dec/01/nestle-robots-coffee-machines-japan-george-clooney-pepper-android-softbank** (accessed September 2018).

Gandhi, A. (2013, February 23). "Basics of Robotics." Slideshare. **https://www.slideshare.net/AmeyaGandhi/basics-of-robotics** (accessed September 2018).

Goris, K., et al. (2010, September). "Mechanical Design of the Huggable Robot Probo." Robotics & Multibody Mechanics Research Group. Brussels, Belgium: Vrije Universiteit Brussels.

Green, D. (2018). "Adidas Just Opened a Futuristic New Factory – and It Will Dramatically Change How Shoes Are Sold." Business Insider. **http://www.businessinsider.com/adidas-high-tech-speedfactory-begins-production-2018-4** (accessed September 2018).

Hernandez, D. (2018). "Seven Jobs Robots Will Create – or Expand." *The Wall Street Journal*. **https://www.wsj.com/articles/seven-jobs-robots-will-createor-expand-1525054021** (accessed September 2018).

History of Robots. (n.d.). Wikipedia. **https://en.wikipedia.org/wiki/History_of_robots** (accessed September 2018).

"Huggable Robot Befriends Girl in Hospital." YouTube video. **https://youtu.be/UaRCCA2rRR0** (accessed August 2018).

"Innovative Human-Robot Cooperation in BMW Group Production." (2013, October 9). BMW Press Release. **https://www.press.bmwgroup.com/global/article/detail/T0209722EN/innovative-human-robot-cooperation-in-bmw-group-production?language=en** (accessed September 2018).

Javelosa, J., & K. Houser. (2017). "Production Soars for Chinese Factory Who Replaced 90% of Employees with Robots." Future Society. **https://futurism.com/2-production-soars-for-chinese-factory-who-replaced-90-of-employees-with-robots/** (accessed September 2018).

Jeong, S., et al. (2015). "A Social Robot to Mitigate Stress, Anxiety, and Pain in Hospital Pediatric Care." *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*.

Jeong, S., et al. (2015). "Designing a Socially Assistive Robot for Pediatric Care." *Proceedings of the Fourteenth International Conference on Interaction Design and Children*. *ACM*.

Jeong, S., & D. Logan. (2018, April 21–26). "Huggable: The Impact of Embodiment on Promoting Socio-emotional Interactions for Young Pediatric Surgeons." MIT Media Lab, Cambridge, MA, CHI 2018, Montréal, QC, Canada.

Jiji. (2017, November 21). "SoftBank Upgrades Humanoid Robot Pepper." *The Japan Times*. **https://www.japantimes.co.jp/news/2017/11/21/business/tech/softbank-upgrades-humanoid-robot-pepper/#.W6B3qPZFzIV** (accessed September 2018).

Joshua, J. (2013, February 24). "The 3 Types of Robots." Prezi. **https://prezi.com/iifjw387ebum/the-3-types-of-robots/** (accessed September 2018).

Kelly, M. (2018, July 16). "This Adorable Robot Wants to Make Air Travel Less Stressful." The Verge. **https://www.theverge.com/2018/7/16/17576334/klm-royal-dutch-airlines-robot-travel-airport** (accessed September 2018).

Kelly, S. M. (2017, August 10). "A Robotic Crib Rocked My Baby to Sleep for Months." CNN Tech. **https://money.cnn.com/2017/08/10/technology/gadgets/snoo-review/index.html** (accessed September 2018).

Lee, K. F. (2018). "The Human Promise of the AI Revolution." *The Wall Street Journal*. **https://www.wsj.com/articles/the-human-promise-of-the-ai-revolution-1536935115** (accessed September 2018).

Mayank. (2012, June 18). "Basic Parts of a Robot." **maxEmbedded.com**. **http://maxembedded.com/2012/06/basic-parts-of-a-robot/** (accessed September 2018).

McHugh, R., & J. Rascon. (2015, May 23). "Meet MEDi, the Robot Taking Pain Out of Kids' Hospital Visits." NBC News. **https://www.nbcnews.com/news/us-news/meet-medi-robot-taking-pain-out-kids-hospital-visits-n363191** (accessed September 2018).

Meister, J. (2017), "The Future Of Work: How Artificial Intelligence Will Transform The Employee Experience," **https://www.forbes.com/sites/jeannemeister/2017/11/09/the-future-of-work-how-artificial-intelligence-will-transform-the-employee-experience/** (accessed November 2018).

Modular Advanced Armed Robotic System. (n.d.). Wikipedia. **https://en.wikipedia.org/wiki/Modular_Advanced_Armed_Robotic_System** (accessed September 2018).

Nagato, T., H. Shibuya, H. Okamoto, & T. Koezuka. (2017, July). "Machine Learning Technology Applied to Production Lines: Image Recognition System." *Fujitsu Scientific & Technical Journal, 53*(4).

O'Kane, S. (2018, May 17). "Raden Is the Second Startup to Bite the Dust After Airlines Ban Some Smart Luggage." Circuit Breaker. **https://www.theverge.com/circuitbreaker/2018/5/17/17364922/raden-smart-luggage-airline-ban-bluesmart** (accessed September 2018).

Park, H. W., et al. (2017). "Growing Growth Mindset with a Social Robot Peer." *Proceedings of the Twelfth ACM/IEEE International Conference on Human Robot Interaction*.

Personal Robots Group. (2016). **https://www.youtube.com/watch?v=sF0tRCqvyT0** (accessed March 5, 2017).

Personal Robots Group, MIT Media Lab. (2017). "Growing Growth Mindset with a Social Robot Peer." *Proceedings of the Twelfth ACM/IEEE International Conference on Human Robot Interaction*.

Prasad, C. (2018, January 22). "Fabio, the Pepper Robot, Fired for 'Incompetence' at Edinburgh Store." *IBN Times*. **https://www.ibtimes.com/fabio-pepper-robot-fired-incompetence-edinburgh-store-2643653** (accessed September 2018).

Ro, L. (2016, October 18). "World's First Smart Crib SNOO Will Help Put Babies to Sleep." Curbed. **https://www.curbed.com/2016/10/18/13322582/snoo-smart-crib-yves-behar-dr-harvey-karp-happiest-baby** (accessed September 2018).

"Robotics Facts." Idaho Public Television. **http://idahoptv.org/sciencetrek/topics/robots/facts.cfm** (accessed September 2018).

"Robots in Agriculture." (2015, July 6). Intorobotics. **https://www.intorobotics.com/35-robots-in-agriculture/** (accessed September 2018).

"Robotics: Types of Robots." **ElectronicsTeacher.com**. **http://www.electronicsteacher.com/robotics/type-of-robots.php** (accessed September 2018).

Rosencrance, L. (2018 May 31). "Tabletop Grapes to Get Picked by Robots in India, with Help from Virginia Tech." Robotics Business Review. **https://www.roboticsbusinessreview.com/agriculture/tabletop-grapes-picked-robots-india-virginia-tech/** (accessed September 2018).

Schuster, W. M. (2018). "Artificial Intelligence and Patent Ownership." *Washington & Lee L. Rev., 75*.

Shadbolt, P. (2015, February 15). "U.S. Navy Unveils Robotic Firefighter." CNN. **https://www.cnn.com/2015/02/12/tech/mci-saffir-robot/index.html** (accessed September 2018).

"Shipboard Autonomous Firefighting Robot – SAFFiR." (2015, February 4). YouTube. **https://www.youtube.com/watch?time_continue=252&v=K4OtS534oYU** (accessed September 2018).

Simon, M. (2018, May 17). "The Wired Guide to Robots." Wired. **https://www.wired.com/story/wired-guide-to-robots/** (accessed September 2018).

"Tabletop Grapes to Get Picked by Robots in India." **Agtechnews.com**. **http://agtechnews.com/Ag-Robotics-Technology/Tabletop-Grapes-to-Get-Picked-by-Robots-in-India.html** (accessed September 2018).

"The da Vinci® Surgical System." (2015, September). Da Vinci Surgery. **http://www.davincisurgery.com/da-vinci-surgery/da-vinci-surgical-system/** (accessed September 2018).

"Types of Robots." (2018). RoverRanch. **https://prime.jsc.nasa.gov/ROV/types.html** (accessed September 2018).

Westlund, J. K., J. M. Lee, J. Plummer, L. Faridia, F. Gray, J. Berlin, M. Quintus-Bosz, H. Harmann, R. Hess, M. Dyer, S. dos Santos, K. Adalgeirsson, S. Gordon, G. Spaulding, S. Martinez, M. Das, M. Archie, M. Jeong, & C. Breazeal, C. (2016). "Tega: A Social Robot." Video Presentation. *Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction*.

White, T. (2015, February 4). "Making Sailors 'SAFFiR' – Navy Unveils Firefighting Robot Prototype at Naval Tech EXPO." America's Navy. **https://www.navy.mil/submit/display.asp?story_id=85459** (accessed September 2018).

Zimberoff, L. (2018, June 21). "A Burger Joint Where Robots Make Your Food." **https://www.wsj.com/articles/a-burger-joint-where-robots-make-your-food-1529599213** (accessed September 2018).

CHAPTER

11

Group Decision Making, Collaborative Systems, and AI Support

LEARNING OBJECTIVES

- Understand the basic concepts and processes of group work, communication, and collaboration
- Describe how computer systems facilitate team communication and collaboration in an enterprise
- Explain the concepts and importance of the time/place framework
- Explain the underlying principles and capabilities of groupware, such as group support systems (GSS)
- Understand how the Web enables collaborative computing and group support of virtual meetings
- Describe collective intelligence and its role in decision making
- Define *crowdsourcing* and explain how it supports decision making and problem solving
- Describe the role of AI in supporting collaboration, group work, and decision making
- Describe human–machine collaboration
- Explain how teams of robots work

In this chapter, we present several topics related to group decision support and collaboration. People work together, and groups (or teams) make many of the complex decisions in organizations. The increase in organizational decision-making complexity drives the need for meetings and group work. Supporting group work in which team members may be in different locations and working at different times emphasizes the important aspects of communications, computer-mediated collaboration, and workplace methodologies. Group support is a critical aspect of decision support systems (DSS). Effective computer-supported group support systems have evolved to increase gains and decrease losses in task performance and underlying processes. New tools and methodology are used to support teamwork. These include collective intelligence, crowdsourcing, and different types of AI. Finally, human–machine and machine–machine collaboration

are increasing the power of collaboration and problem solving. All these are presented in the following sections:

11.1 OPENING VIGNETTE: Hendrick Motorsports Excels with Collaborative Teams

Hendrick Motorsports (HMS) is a leading car racing company (with more than 500 employees) that competes in the Monster Energy NASCAR Cup Series. HMS's major objective is to win as many races as possible each year. The company enters four race cars and their teams. HMS also builds its race cars. This includes building or rebuilding 550 car engines every year. In this kind of business, teamwork is critical because many different people with different skills and knowledge and several professional teams contribute to the success of the company.

THE OPERATIONS

HMS is engaged in car races all over the United States during the racing season (38 weeks a year). The company moves to a different racetrack every week. During the off-season time (14 weeks), the company analyzes the data obtained, and lessons learned during the latest racing seasons, and prepares for the following season. The company's headquarters contains 19 buildings scattered over 100 acres.

THE PROBLEMS DURING THE RACING SEASON

The company needs to make quick decisions during races—some in real time, sometimes in a split second. Different team members need to participate, and they are in different locations. Communication and collaboration are critical.

Car racing is based on teamwork, drivers, engineers, planners, mechanics, and others who participate. Members must communicate and collaborate to make decisions.

The environment is too noisy to talk during a race. However, team members need to share data, graphs, and images, and chat quickly. Several decisions need to be made in real time that will help win races (e.g., how much fuel to add in the next few seconds to a car in the middle of the race). Team members must communicate and share data, including visual. It takes about 45–50 seconds for a car to complete a 2.5-mile lap at Daytona 500. During the race, top engineers need to communicate constantly with the fuelers. Last-minute data are common during the racing session.

Any knowledge acquired in each lap can be used to improve the next one. In races, fueling decisions are critical. There are many other decisions to be made during the racing season. For example, after each race, the company needs to move a large crew with equipment and supplies from one location to the next (38 different venues). Moves need to be fast, efficient, and economical. Again, teamwork, as well as coordination, is needed.

OFF-SEASON PROBLEMS

There are 14 weeks to prepare for the next season. In addition, there is a considerable amount of data to analyze, simulate, discuss, and manipulate. For this, people need not only communication and collaboration tools but also analytics of different types.

THE SOLUTION

HMS decided to use Microsoft Teams, which is a chat-based platform, for team workspace in Microsoft Office 365. This platform is used as a *communication hub* for team members at the race tracks and at any other location in the organization.

Microsoft Teams stores data in different formats in its Teams workspace. Therefore, car crews, engineers, and mechanics can make split-second decisions that may help win races. This also enables computational analysis in a central place.

Microsoft Teams includes several subprograms and is easily connected to other software in Office 365. Office 365 provides several other tools that increase collaboration (e.g., SharePoint). For example, in the HSM solution, there is a working link to Excel as well as to SharePoint. Also, One Note of Teams is used to share meeting notes. Before Teams, the company used Slack (Section 11.4), but Slack did not provide enough security and functions.

Members need to share and discuss the massive amount of data accumulated during the racing season. Note that several employees have multiple skills and tasks. The solution included the creation of a *collaboration hub* for concurrent projects. Note that each different project may require different talents and data, depending on the project's type. Also, the solution involves other information technology (IT) tools. For example, HMS uses Power BI dashboard to communicate data visually. Some data can be processed as Excel-based spreadsheets.

Microsoft Teams is also available as a mobile app. Each team's data file is available on the track at home and even under a car. So, the software package is able to respond to important situations right away.

The Results

The major results were improved productivity, smoother communication, easier collaboration, and reduction of the need for the time consumed in face-to-face meetings. People can chat online, seeing their partners without leaving their physical workplace. The company admits that without Teams, it would not have been able to accomplish its success. Today, Teams has everything the company needs at its fingertips.

QUESTIONS FOR THE OPENING VIGNETTE

1. What were the major drivers for the use of Microsoft's Teams?
2. List some discussions held during the racing season, and how they were supported by the technology.
3. List decisions held during the off-season, and how they were supported by the technology.
4. Discuss why Microsoft Teams was selected, and explain how it supports teamwork group decision making.
5. Trace communication and collaboration within and between groups.

6. Specify the function of Microsoft Teams workspace.
7. Watch the video at **youtube.com/watch?time_continue=108&v=xnFdM9IOaTE** and summarize its content.

WHAT WE CAN LEARN FROM THIS VIGNETTE

The first lesson is that many tasks today must be done by collaborating teams in order to succeed. Second, time is critical; therefore, companies must use technology to speed operations and facilitate communication and collaboration in teamwork. Third, it is possible to use existing software for support, but it is better to use a major vendor that has additional products that can supplement the collaboration/communication software. Fourth, chatting can expedite communication, and visual technology support can be useful. Fifth, team members belong to diverse units and have diverse skills. The software brings them together. Team members should have clear goals and understand how to achieve them. Finally, collaboration can be both within and between groups.

Sources: Compiled from Ruiz-Hopper (2016) and Microsoft (2017).

11.2 MAKING DECISIONS IN GROUPS: CHARACTERISTICS, PROCESS, BENEFITS, AND DYSFUNCTIONS

Managers and other knowledge workers continuously make decisions, design products, develop policies and strategies, create software systems, and so on. Frequently they do it in groups. When people work in groups (i.e., teams), they perform group work or teamwork. **Group work** refers to work done by two or more people together. One aspect of group work is group decision making.

Group decision making refers to a situation in which people make decisions together. Let's first look at the characteristics of group work.

Characteristics of Group Work

The following are some of the functions and characteristics of group work:

- Group members may be located in different places.
- Group members may work at different times.
- Group members may work for the same organization or different organizations.
- A group can be permanent or temporary.
- A group can be at one managerial level or span several levels.
- A group can create synergy (leading to process and task gains) or result in conflict.
- A group can generate productivity gains and/or losses.
- A group's task may have to be accomplished very quickly.
- It may be impossible or too expensive for all team members to meet in one place at the same time, especially when the meeting is called for emergency purposes.
- Some of the groups' needed data, information, or knowledge may be located in several sources, some of which may be external to the organization.
- The expertise of a group's team members may be needed.
- Groups perform many tasks; however, groups of managers and analysts frequently concentrate on decision making or problem solving.
- The decisions made by a group are easier to implement if supported by all (or at least most) members.
- Group work has many benefits and, unfortunately, some possible dysfunctions.
- Group behaviors are influenced by several factors and may affect group decisions.

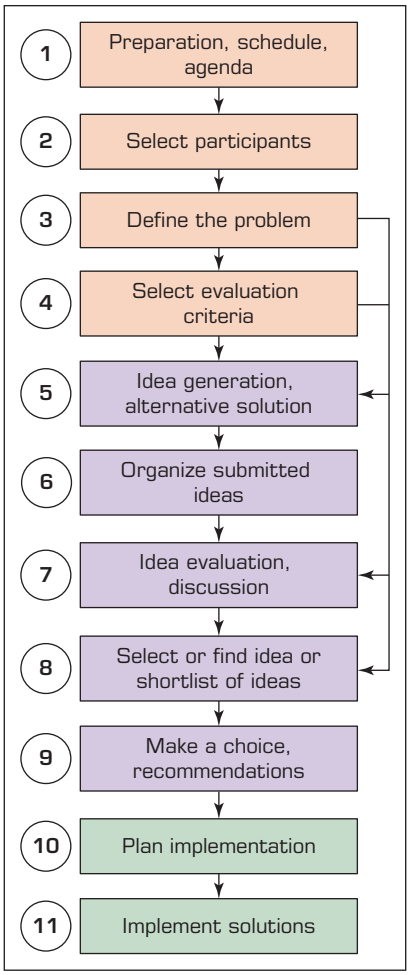


FIGURE 11.1 **The Process of Group Decision Making.**

Types of Decisions Made by Groups

Groups are usually involved in two major types of decision making:

1. Making a decision together.
2. Supporting activities or tasks related to the decision-making process. For example, the group may select criteria for evaluating alternative solutions, prioritizing possible ones, and helping design strategy to implement them.

Group Decision-Making Process

The process of group decision making is similar to that of the general decision-making process described in Chapter 1 but it has more steps. Steps of the group decision-making process are illustrated in Figure 11.1.

Step 1. Prepare for meetings regarding the agenda, time, place, participants, and schedule.
Step 2. Determine the topic of the meeting (e.g., problem definition).
Step 3. Select participants for the meeting.
Step 4. Select criteria for evaluating the alternatives and the selected solution.
Step 5. Generate alternative ideas (brainstorm).
Step 6. Organize the ideas generated into similar groups.
Step 7. Evaluate the ideas, discuss, and brainstorm.

Step 8. Select a short list (finalists).
Step 9. Select a recommended solution.
Step 10. Plan implementation of the solution.
Step 11. Implement the solution.

The process is shown as sequential, but as shown in Figure 11.1, some loops are possible. Also, if no solution is found, the process may start again.

GROUP DECISION FACTS When a group is going through the steps shown in Figure 11.1, the following is usually true:

- The decisions made need to be implemented.
- Group members are typically of equal or nearly equal status.
- The outcome of a meeting depends partly on the knowledge, opinions, and judgments of its participants and the support they give to the outcome.
- The outcome of a meeting depends on the composition of the group and on the decision-making process it uses.
- Group members settle differences of opinions either by the ranking person present or through negotiations or arbitration.
- The members of a group can be in one place, meeting face-to-face, or they can be a **virtual team**, in which case they are in different places meeting electronically. They can also meet at different times.

Benefits and Limitations of Group Work

Some people endure meetings (the most common form of group work) as a necessity; others find meetings to be a waste of time. Many things can go wrong in a meeting. Participants may not clearly understand its purpose, may lack focus, or may have hidden agendas. Many participants may be afraid to speak up, or a few may dominate the discussions. Misunderstandings occur because of different interpretations of language, gestures, or expression. Technology Insight 11.1 provides a list of factors that can hinder the effectiveness of a manually managed meeting. Besides being challenging, teamwork is also expensive. A meeting of several top managers or executives can cost thousands of dollars.

Group work may have potential benefits (process gains) or drawbacks (process losses). **Process gains** are the benefits of working in groups. The unfortunate dysfunctions that may occur when people work in groups are called **process losses**. Examples of each are listed in Technology Insight 11.1.

TECHNOLOGY INSIGHT 11.1 Benefits and Dysfunctions of Working in Groups

The following are the possible major benefits and dysfunctions of group works.

Benefits of Working in Groups (Process Gains)	Dysfunctions of Face-to-Face Group Process (Process Losses)
• It provides learning. Groups are better than individuals at understanding problems. They can teach each other.	• Social pressures of conformity may result in **groupthink** (i.e., people begin to think alike and not tolerate new ideas; they yield to *conformance pressure*).
• People readily take ownership of problems and their solutions.	• It is a time-consuming, slow process. • Some relevant information could be missing.
• Group members have their egos embedded in the final decision, so they are committed it.	• A meeting can lack coordination, have a poor agenda, or be poorly planned.

Benefits of Working in Groups (Process Gains)	Dysfunctions of Face-to-Face Group Process (Process Losses)
• Groups are better than individuals at catching errors.	• A meeting may be dominated by time, topic, opinion of one or a few individuals, or fear of contributing because of the possibility of conflicts.
• A group has more *information* and knowledge than any one member does. Members can combine their knowledge to create new knowledge. More and more creative alternatives for problem solving can be generated, and better solutions can be derived (e.g., through *brainstorming*).	• Some group members can tend to influence the agenda while some try to rely on others to do most of the work (free riding). The group may ignore good solutions, have poorly defined goals, or be composed of the wrong participants.
• A group may produce *synergy* during problem solving, therefore the effectiveness and/or quality of group work can be greater than the sum of what individual members produce.	• Some members may be afraid to speak up. • The group may be unable to reach consensus. • The group may lack focus.
• Working in a group may stimulate the creativity of the participants and the process.	• There can be a tendency to produce poor-quality compromises.
• Working together could allow a group to have better and more precise communication.	• There is often nonproductive time (e.g., socializing, preparing, waiting for latecomers).
• Risk propensity is balanced. Groups moderate high-risk takers and encourage conservatives.	• There can be a tendency to repeat what has already been said (because of failure to remember or process).
	• Meeting costs can be high (e.g., travel, participation time spent).
	• There can be incomplete or inappropriate use of information.
	• There can be too much information (i.e., information overload).
	• There can be incomplete or incorrect task analysis.
	• There can be inappropriate or incomplete representation in the group.
	• There can be attention or concentration blockage.

SECTION 11.2 REVIEW QUESTIONS

1. Define *group work*.
2. List five characteristics of group work.
3. Describe the steps of group decision making.
4. List the major activities that occur in group work.
5. List and discuss five benefits of group work.
6. List and discuss five dysfunctions of group-made decisions.

11.3 SUPPORTING GROUP WORK AND TEAM COLLABORATION WITH COMPUTERIZED SYSTEMS

When people work in teams, especially when the members are in different locations and may work at different times, they need to communicate, collaborate, and access a diverse set of information sources in multiple formats. This makes meetings, especially virtual ones, complex with an increased chance for process losses. Therefore, it is important to follow certain processes and procedures for conducting meetings.

Group work may require different levels of coordination. Sometimes a group operates at the individual work level with members making individual efforts that require

no coordination. As with a team of sprinters representing a country participating in a 100-meter dash, group productivity is simply the best of the individual results. At other times, group members may interact in coordination. At this level, as with a team in a relay race, the work requires careful coordination between otherwise independent individual efforts. Sometimes a team may operate at the concerted work level. As in a rowing race, teams working at this level must make a continuous concerted effort to be successful. Different mechanisms support group work at different levels of coordination.

Most organizations, small and large, use some computer-based communication and collaboration methods and tools to support people working in teams or groups. From e-mails to mobile phones and Short Message Service (SMS), as well as conferencing technologies, such tools are an indispensable part of today's work life. We next highlight some related technologies and applications.

Overview of Group Support Systems (GSS)

For groups to collaborate effectively, appropriate communication methods and technologies are needed. We refer to these technologies as **group support systems (GSS)**. The Internet and its derivatives (i.e., intranets, Internet of Things [IoT], and extranets) are the infrastructures on which much communication and collaboration occurs. The Web supports intra- and inter-organizational collaborative decision making.

Computers have been used for several decades to facilitate group work and decision making. Lately, collaborative tools have received more attention due to their increased capabilities and ability to save time and money (e.g., on travel cost) and to expedite decision making. Computerized tools can be classified according to time and place categories.

Time/Place Framework

The tools used to support collaboration, groups, and the effectiveness of collaborative computing technology depend on the location of the group members and on the time that shared information is sent and received. DeSanctis and Gallupe (1987) proposed a framework for classifying IT communication support technologies. In this framework, communication is divided into four cells, which are shown with representative computerized support technologies in Figure 11.2. The four cells are organized along two dimensions—time and place.

When information is sent and received almost simultaneously, the communication is in **synchronous (real-time)** mode. Telephones, instant messaging (IM), and face-to-face meetings are examples of synchronous communication. **Asynchronous** communication occurs when the receiver gets (or views) the information, such as an e-mail, at a different time than it was sent. The senders and the receivers can be in the same place or in different places.

As shown in Figure 11.2, time and place combinations can be viewed as a four-cell matrix, or framework. The four cells of the framework are as follows:

- ***Same time/same place.*** Participants meet face-to-face, as in a traditional meeting, or decisions are made in a specially equipped decision room. This is still an important way to meet even when Web-based support is used because it is sometimes critical for participants to leave their regular workplace to eliminate distractions.
- ***Same time/different place.*** Participants are in different places, but they communicate at the same time (e.g., with videoconferencing or IM).
- ***Different time/same place.*** People work in shifts. One shift leaves information for the next shift.
- ***Different time/(any place) different place (any place).*** Participants are in different places, and they send and receive information at different times. This occurs when team members are traveling, have conflicting schedules, or work in different time zones.

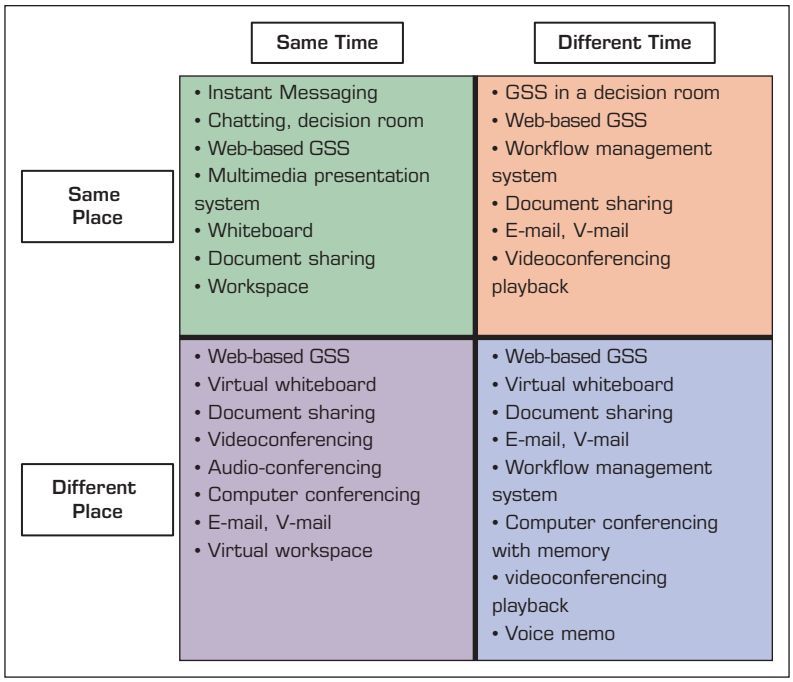


FIGURE 11.2 **The Time/Place Framework.**

Groups and group work in organizations are proliferating. Consequently, groupware continues to evolve to support effective group work, mostly for communication and collaboration (Section 11.4).

Group Collaboration for Decision Support

In addition to making decisions, groups also support decision-making subprocesses such as brainstorming. Collaboration technology is known to be the driving force for productivity increase and boosting people and organizational performance. Groups collaborate to make decisions in several ways. For example, groups provide assistance for the steps in Figure 11.1. Groups can help to identify problems, to assist in choosing criteria for selecting solutions, generating solutions (e.g., brainstorming), evaluating alternatives, and assisting in the selection of the best solution and implementing it. The group can be involved in one step or in several steps. In addition, it can collect the necessary data.

Many technologies can be used for collaboration; several of them are computerized and are described in several sections in this chapter.

Studies indicate that adopting collaboration technologies increases productivity: for example, visual collaborative solutions increase employees' satisfaction and productivity.

COMPUTERIZED TOOLS AND PLATFORMS We divide the computerized support into two parts. In Section 11.4, we present the major support of generic activities in communication and collaboration. Note that hundreds, maybe thousands, of commercial products are available to support communication and collaboration. We cover only a sample here.

Section 11.5 covers direct support of decision making, both to the entire process and to the major steps in the process. Note that some products, such as Microsoft Teams, which is cited in the opening vignette, support both generic activities and those in the decision-making process.

SECTION 11.3 REVIEW QUESTIONS

1. Why do companies use computers to support group work?
2. What is GSS?
3. Describe the components of the time/place framework.
4. Describe the importance of collaboration for decision making.

11.4 ELECTRONIC SUPPORT FOR GROUP COMMUNICATION AND COLLABORATION

A large number of tools and methods are available to facilitate group work, e-collaboration, and communication. The following sections present only some tools that support the process. Our attention here is on indirect support to decision making. In Section 11.5, we cover direct support.

Groupware for Group Collaboration

Many computerized tools have been developed to provide group support. These tools are called **groupware** because their primary objective is to support group work indirectly as described in this section. Some e-mail programs, chat rooms, IM, and teleconferences provide indirect support.

Groupware provides a mechanism for team members to share opinions, data, information, knowledge, and other resources. Different computing technologies support group work in different ways depending on the task and size of the group, the security required, and other factors.

CATEGORIES OF GROUPWARE PRODUCTS AND FEATURES Many groupware products to enhance the collaboration of a small and large number of people are available on the Internet or intranets. A prime example is Microsoft's Teams (opening vignette). The features of groupware products that support commutation, collaboration, and coordination are listed in Table 11.1. What follows are brief definitions of some of those features.

Synchronous versus Asynchronous Products

The products and features described in Table 11.1 may be synchronous or asynchronous. Web conferencing and IM, as well as voice-over IP (VoIP), are associated with the synchronous mode. Methods associated with asynchronous modes include e-mail and **online workspaces** where participants can collaborate while working at different times. Google Drive (**drive.google.com**) and Microsoft SharePoint (**http://office.microsoft.com/en-us/SharePoint/collaboration-software-SharePoint-FX103479517.aspx**) allow users to set up online workspaces for storing, sharing, and working collaboratively on different types of documents. Similar products are Google Cloud Platform and Citrix Workspace Cloud.

Companies such as **Dropbox.com** provide an easy way to share documents. Similar systems, such as photo sharing (e.g., Instagram, WhatsApp, Facebook), are evolving for consumer home use.

TABLE 11.1 Groupware Products and Features

General (Can Be Either Synchronous or Asynchronous)
• Built-in e-mail, messaging system
• Browser interface
• Joint Web page creation
• Active hyperlink sharing
• File sharing (graphics, video, audio, or other)
• Built-in search functions (by topic or keyword)
• Workflow tools
• Corporate portals for communication, collaboration, and search
• Shared screens
• Electronic decision rooms
• Peer-to-peer networks
Synchronous (Same Time)
• IM
• Videoconferences, multimedia conferences
• Audioconferences
• Shared whiteboard, smart whiteboard
• Instant videos
• Brainstorming
• Polling (voting) and other decision support (activities such as consensus building, scheduling)
• Chats with people
• Chats with bots
Asynchronous (Different Times)
• Virtual workspaces
• Tweets
• Ability to receive/send e-mail, SMS
• Ability to receive notification alerts via e-mail or SMS
• Ability to collapse/expand discussion threads
• Message sorting (by date, author, or read/unread)
• Auto responders
• Chat session logs
• Electronic bulletin boards, discussion groups
• Blogs and wikis
• Collaborative planning and/or design tools

Groupware products are either stand-alone, supporting one task (such as videoconferencing), or integrated, including several tools. In general, groupware technology products are fairly inexpensive and can easily be incorporated into existing information systems.

Virtual Meeting Systems

The advancement of Web-based systems opens the door for improved electronically supported **virtual meetings** with the virtual team members in different locations, even in different countries. Online meetings and presentation tools are provided by tools such as *webex*, **GoToMeeting.com**, **Skype.com**, and many others. These systems feature Web

seminars (popularly called Webinars), screen sharing, audioconferencing, videoconferencing, polling, question–answer sessions, and so on. Microsoft Office 365 includes a built-in virtual meeting capability. Even smartphones now have sufficient interaction capabilities to allow live meetings through applications such as FaceTime.

COLLABORATIVE WORKFLOW *Collaborative workflow* refers to software products that address project-oriented and collaborative processes. They are administered centrally yet are capable of being accessed and used by workers from different departments and from different physical locations. The goal of collaborative workflow tools is to empower knowledge workers. The focus of an enterprise solution for collaborative workflow is on allowing workers to communicate, negotiate, and collaborate within an integrated environment. Some leading vendors of collaborative workflow applications are FileNet and Action Technologies. Collaborative workflow is related to but different than collaborative workspace.

DIGITAL COLLABORATIVE WORKSPACE: PHYSICAL AND VIRTUAL A **collaborative workspace** is where people can work together from any location at the same or at a different time. Originally, it was a physical conference room that teams used for conducting meetings. It was expanded to be a shared workspace, also known as "coworking space." Some of these are in companies; others are offered for rent. Different computerized technologies are available to support group work in a physical structure. For 12 benefits of collaborative workspace, see Pena (2017).

A *virtual collaboration workspace* is an environment equipped with digital support by which group members who are in different locations can share information and collaborate. A simple example is Google Drive, which enables sharing spreadsheets.

Collaborative workspace enables tech-savvy employees to access systems and tools from any device they need. People can work together in a secure way from anywhere. The digital workspace increases team productivity and innovation. It empowers employees and unlocks innovation. It allows workers to reach other people for collaborative work. For details and other collaboration technologies, see de Lares Norris (2018).

Example

PricewaterhouseCoopers (PwC) built an ideation war room in its Paris office as a large, immersive collaboration facility to support customer meetings.

MAJOR VENDORS OF VIRTUAL WORKSPACE Products by five major vendors follow:

- *Google Cloud Platform* is deployed on the "cloud," so it is offered as a platform-as-a-service (PaaS). Google is also known for its Flexible Workspace product.
- *Citrix Workspace Cloud* is also deployed on the "cloud." Citrix is known for its GoToMeeting collaboration tool. *Citrix Workspace Cloud* users can manage secure digital workplaces on Google Cloud.
- *Microsoft Workspace* is part of Office 365.
- *Cisco's Webex,* a popular collaboration package including Meeting.
- *Slack workspace* is a very popular workspace.

ESSENTIALS OF SLACK Slack workspace is a digital space on which teammates share, communicate, and collaborate on work. It can be in one organization, or large organizations may have multiple interconnected Slack spaces.

Each workspace includes several topical channels. These can be organized as public, private, or shared. The remaining components of Slack are messages, searches, and notifications. There are four groups of people involved with Slack: workspace owners,

workspace administrators, members, and guests. For a Slack Guide, see **get.slack.help/hc/en-us/articles/115004071768-What-is-Slack-**.

Slack has many key features and can deliver secure virtual apps to almost any device.

Collaborative Networks and Hubs

Traditionally, collaboration has taken place among supply chain members, frequently those that were close to each other (e.g., a manufacturer and its distributor or a distributor and a retailer). Even if more partners were involved, the focus was on the optimization of information and product flow between existing nodes in the traditional supply chain. Advanced methods, such as collaborative planning, forecasting, and replenishment (CPFR), do not change this basic structure.

Traditional collaboration results in a vertically integrated supply chain. However, Web technologies can fundamentally change the shape of the supply chain, the number of players in it, and their individual roles. In a collaborative network, partners at any point in the network can interact with each other, bypassing what are traditional partners. Interaction may occur among several manufacturers or distributors as well as with new players, such as software agents that act as aggregators.

Collaborative Hubs

The purpose of a collaborative hub is to be a center point for group collaboration.

Collaborative hub platforms need to enable participants' interactions to unfold in various forms online.

Example: Surface Hub for Business by Microsoft

This product connects individuals wherever they are and whenever they want to use a digital whiteboard and integrating software and apps. It helps to create a collaboration workplace where multiple devices are connected wirelessly to create a powerful work environment.

Social Collaboration

Social collaboration refers to collaboration conducted within and between socially oriented groups. It is a process of group interactions and information/knowledge sharing while attempting to attain common goals. Social collaboration is usually done on social media sites, and it is enabled by the Internet, IoT, and diversified social collaboration software. Social collaboration groups and schemes can take many different shapes. For images, conduct a Google search for "images of social collaboration."

COLLABORATION IN SOCIAL NETWORKS Business-related collaboration is most evidenced on Facebook and LinkedIn. However, Instagram, Pinterest, and Twitter support collaboration as well.

- *Facebook.* Facebook's Workspace **facebook.com/workspace** is used by hundreds of thousands of companies utilizing its features, such as "groups," to support team members. For example, 80 percent of Starbucks store managers use this software.
- *LinkedIn.* LinkedIn provides several collaboration tools to its members. For example, LinkedIn Lookup provides several tools. Also, LinkedIn is a Microsoft company and it provides some integrated tools. The creation of subgroups of interest is a useful facilitator.

SOCIAL COLLABORATION SOFTWARE FOR TEAMS In addition to the generic collaboration software that can be used by two people and by teams, there are software platforms specifically for forming teams and supporting their activities. A few popular examples

according to **collaboration-software.financesonline.com/c/social-collaboration-software/** are Wrike, Ryver, Azendoo, Zimbra social platform, Samepage, Zoho, Asana, Jive, Chatter, and Social Tables. For viewing the best social collaboration software by category, see **technologyadvice.com/social-collaboration-software/**.

Sample of Popular Collaboration Software

As noted earlier, there are hundreds or maybe thousands of communication and collaboration software products. Furthermore, their capabilities are ever changing. Given that our major interest is decision-making support, we provide only a small sample of these tools. We use the classification and example of Time Doctor, using the 2018 list (see Digneo, 2018).

- Communication tools: Yammer (social collaboration), Slack, Skype, Google Hangouts, GoToMeeting
- Design tools: InVision, Mural, Red Pen, Logo Maker
- Documentation tools: Office Online, Google Docs, Zoho
- File-sharing tools: Google Drive, Dropbox, Box
- Project management tools: Asana, Podio, Trello, WorkflowMax, Kanban Tool,
- Software tools: GitHub, Usersnap,Workflow tools: Integrity, BP Logix

OTHER TOOLS THAT SUPPORT COLLABORATION AND/OR COMMUNICATION

Notejoy (makes collaborative notes for team).
Kahootz (brings stakeholders together to form communities of interest).
Nowbridge (offers team connectivity, ability to see participants).
Walkabout Workplace (is a 3D virtual office for remote teams).
RealtimeBoard (is a enterprise visual collaboration).
Quora (is a popular place for posting questions to the crowd).
Pinterest (provides an e-commerce workspace that allows collection of text and images on selected topics).
IBM connection closed (offers a comprehensive communication and collaboration tool set).
Skedda (schedules space for coworking)
Zinc (is a social collaboration tool)
Scribblar (is an online collaboration room for virtual brainstorming)
Collokia (is a machine learning platform for workflow)
For additional tools, see Steward (2017).

SECTION 11.4 REVIEW QUESTIONS

1. Define *groupware*.
2. List the major groupware tools and divide them into synchronous and asynchronous types.
3. Identify specific tools for Web conferencing and their capabilities.
4. Describe collaborative workflow.
5. What is collaborative workspace? What are its benefits?
6. Describe social collaboration.

11.5 DIRECT COMPUTERIZED SUPPORT FOR GROUP DECISION MAKING

Decisions are made frequently at meetings, some of which are called in order to make a one-time specific decision. For example, directors are elected at shareholder meetings, organizations allocate budgets in meetings, cities decide which candidates to hire for their top

positions, and the U.S. federal government meets periodically to set the short-term interest rate. Some of these decisions are complex; others can be controversial, as in resource allocation by a city government. Process dysfunctions can be significantly large in such situations; therefore, computerized support has often been suggested to mitigate these controversies. These computer-based support systems have appeared in the literature under different names, including *group decision support systems* (GDSS), *group support systems* (GSS), *computer-supported collaborative work* (CSCW), and *electronic meeting systems* (EMS). These systems are the subject of this section. In addition to supporting entire processes, there are tools that support one or several activities in the group decision-making process (e.g., brainstorming).

Group Decision Support Systems (GDSS)

During the 1980s, researchers realized that computerized support to managerial decision making needed to be expanded to groups, because major organizational decisions are made by groups, such as executive committees and special task forces. The result was the creation of the *group decision support systems* methodology.

A **group decision support system (GDSS)** is an interactive computer-based system that facilitates the solution of semistructured or unstructured problems by a group of decision makers. The goals of GDSS are to improve the productivity of decision-making meetings by speeding up the decision-making *process* and/or to increase the quality of the resulting decisions.

MAJOR CHARACTERISTICS AND CAPABILITIES OF A GDSS GDSS characteristics follow:

- It supports the *process* of group decision makers mainly by providing automation of subprocesses (e.g., brainstorming) and using information technology tools.
- It is a specially designed information system, not merely a configuration of already existing system components. It can be designed to address one type of problem or make a variety of group-level organizational decisions.
- It encourages generation of ideas, resolution of conflicts, and freedom of expression. It contains built-in mechanisms that discourage development of negative group behaviors, such as destructive conflict, miscommunication, and groupthink.

The first generation of GDSS was designed to support face-to-face meetings in a *decision room*. Today, support is provided mostly over the Web to virtual teams. A group can meet at the same time or at different times. GDSS is especially useful when controversial decisions have to be made (e.g., resource allocation, determining which individuals to lay off). GDSS applications require a facilitator for one physical place or a coordinator or leader for online virtual meetings.

GDSS can improve the decision-making process in various ways. For one, GDSS generally provides structure to the meeting planning process, which keeps a group meeting on track, although some applications permit the group to use unstructured techniques and methods for **idea generation**. In addition, GDSS offers rapid and easy access to external and stored information needed for decision making. It also supports parallel processing of information and idea generation by participants and allows asynchronous computer discussion. GDSS makes possible larger group meetings that would otherwise be unmanageable; having a larger group means that more complete information, knowledge, and skills can be represented in the meeting. Finally, voting can be anonymous with instant results, and all information that passes through the system can be recorded for future analysis (producing *organizational memory*).

Over time, it became clear that supporting teams needed to be broader than GDSS has beed supported in a decision room. Furthermore, it became clear that what was really

needed was support for *virtual teams*, both in different place/same time and different place/different time situations. Also, it became clear that teams needed indirect support in most decision-making cases (e.g., help in searching for information or in collaboration) rather than direct support for the decision-making process. Although GDSS expanded to virtual team support, it was unable to meet all the other needs. In addition, the traditional GDSS was designed to deal with contradictory decisions when conflicts were likely to arise. Thus, a new generation of GDSS that supports collaboration work was needed. As we will see later, products such as Stormboard provide those needs.

Characteristics of GDSS

There are two options for deploying GDSS technology: (1) in a special-purpose decision room and (2) as Internet-based groupware with client programs running wherever the group members are.

DECISION ROOMS The earliest GDSS was installed in expensive, customized, special-purpose facilities called **decision rooms** (or electronic meeting rooms) that had PCs and a large public screen at the front of each room. The original idea was that only executives and high-level managers would use the expensive facility. The software in an electronic meeting room usually ran over a local area network (LAN), and these rooms were fairly plush in their furnishings. Electronic meeting rooms were structured in different shapes and sizes. A common design was a room equipped with 12 to 30 networked PCs, usually recessed into the desktop (for better participant viewing). A server PC was attached to a large screen projection system and connected to the network to display the work at individual workstations and aggregated information from the facilitator's workstation. Breakout rooms equipped with PCs connected to the server, in which small subgroups could consult, were sometimes located adjacent to the decision room. The output from the subgroups was able to be displayed on the large public screen. A few companies offered such rooms for a daily rent. Only a few upgraded rooms are still available today, usually for high rent.

INTERNET-BASED GROUPWARE Since the late 1990s, the most common approach to GSS and GDSS delivery has been to use an Internet-based groupware that allows group members to work from any location at any time (e.g., WebEx, GoToMeeting, Adobe Connect, IBM Connections, Microsoft Teams). This groupware often includes audio conferencing and videoconferencing. The availability of relatively inexpensive groupware, as described in Section 11.4, combined with the power and low cost of computers and the availability of mobile devices, makes this type of system very attractive.

Supporting the Entire Decision-Making Process

The process that was illustrated in Figure 11.1 can be supported by a variety of software products. In this section, we provide an example of one product, Stormboard, that supports several aspects of that process.

Example: Stormboard

Stormboard **stormboard.com** provides support for different brainstorming and group decision-making configurations. The following is the product's sequence of activities:

1. Define the problem and the users' objectives (what they are hoping to achieve).
2. Brainstorm ideas (to be discussed later).
3. Organize the ideas in groups of similar flavor, look for patterns, and select only viable ideas.

4. Collaborate, refine concepts, and evaluate (using criteria) the meeting's objectives.
5. The software enables users to prioritize proposed ideas by focusing on the selection criteria. It lets all participants express their thinking and directs the team to be cohesive.
6. It presents a short list of superior ideas.
7. The software suggests the best idea and recommends implementation.
8. It plans the project implementation.
9. It manages the project.
10. It periodically reviews progress.

For a video, see **youtube.com/watch?v=0buRzu4rhJs**.

COMPREHENSIVE GROUPWARE TOOLS INCLUDING THINKTHANK Although many capabilities that enable group decision support are embedded in common software tools for office productivity such as Microsoft Office 365, it is instructive to learn about specific software that illustrates some of groupware's unique capabilities. MeetingRoom was one of the first comprehensive, same time/same place electronic meeting packages. Its follow-up product, GroupSystems OnLine, offered similar capabilities, and it ran in asynchronous mode (anytime/anyplace) over the Web (MeetingRoom ran only over a LAN). GroupSystems' latest product is ThinkTank, a suite of tools that facilitate the various group decision-making activities. For example, it shortens cycle time for brainstorming. ThinkTank improves the collaboration of face-to-face or virtual teams through customizable processes toward the groups' goals faster and more effectively than in previous product generations. ThinkTank offers the following:

- It can provide efficient participation, workflow, prioritization, and decision analysis.
- Its anonymous brainstorming for ideas and comment generation is an ideal way to capture the participants' creativity and experience.
- The product's enhanced Web 2.0 user interface ensures that participants do not need special training to join, so they can focus 100 percent on solving problems and making decisions.
- With ThinkTank, all of the knowledge shared by participants is captured and saved in documents and spreadsheets, automatically converted to the meeting minutes, and made available to all participants at the end of the session.

Examples: ThinkTank Use (thinktank.net/case-study)

The following are two examples of ThinkTank's use.

- It enables transformational collaboration between supply chain partners. Their meeting was supported by collective intelligence tools and procedures. Partners agreed on how to cut costs, speed processes, and improve efficiencies. In the past, there had been no progress on these issues.
- The University of Nebraska and the American College of Cardiology collaborated using ThinkTank tools and procedures to rethink how electronic health records could be reorganized to help medical consultants save time. Patients' appointment times were shortened by 5 to 8 minutes. Other improvements also were achieved. Both patient care and large monetary savings were achieved.

OTHER DECISION-MAKING SUPPORT The following is a list of other types of support provided by intelligent systems:

- Using knowledge systems and a product called Expert Choice Software for dealing with multiple-criteria group decision making.

- A mediating group decision-making method for infrastructure asset management was proposed by Yoon et al. (2017).
- For a group decision-making support system in logistics and supply chain management, see Yazdani et al. (2017).

Brainstorming for Idea Generation and Problem Solving

A major activity in group decision making is idea generation. **Brainstorming** is a process for generating creative ideas. It involves freewheeling group discussions and spontaneous contribution of ideas for solving problems and making strategy and resource allocation. Contributors' ideas are discussed by the members. An attempt is made to generate as many ideas as possible, no matter how bizarre they look. Generated ideas are discussed and evaluated by the group. There is evidence that groups not only generate more ideas but also better ones (McMahon et al., 2016). Manually managed brainstorming has some of the limitations of group work described in Section 11.2. Therefore, computer support is frequently recommended.

COMPUTER-SUPPORTED BRAINSTORMING Computer programs can support the various brainstorming activities. The support is usually for online brainstorming, synchronously or asynchronously. Hopefully, electronic brainstorming eliminates many of the process dysfunctions cited in Section 11.2 and helps in the generation of many new ideas. Brainstorming software can stand alone or be a part of a general group support package. The major features of software packages follow:

- Creation of a large number of ideas.
- Large group participation.
- Real-time updates.
- Information color coding.
- Collaborative editing.
- Design of brainstorming sessions.
- Idea sharing.
- People participation.
- Idea mapping (e.g., create mind maps).
- Text, video, documents, etc. posting.
- Remote brainstorming.
- Creation of an electronic archive.
- Reduction of social loafing.

The major limitations of electronic software support are increased cognitive load, fear of using new technology, and need for technical assistance.

COMPANIES THAT PROVIDE ONLINE BRAINSTORMING SERVICES AND SUPPORT FOR GROUP WORK Some companies and the services and support they provide follow:

- **eZ Talks Meetings.** Cloud-based tool for brainstorming and idea sharing.
- **Bubbl.us.** Visual thinking machine that provides a graphical representation of ideas and concepts, helps in idea generation, and shows where ideas and thoughts overlap (visually, in colors).
- **Mindomo.** Tool for real-time collaboration that offers integrated chat capability.
- **Mural.** Tool that enables collecting and sorting of ideas in rich media files. It is designed as a Pinboard that invites participants.
- **iMindQ.** Cloud-based service that enables creating mind maps and basic diagrams.

For an evaluation of 28 online brainstorming tools, see **blog.lucidmeetings.com/blog/25-tools-for-online-brainstorming-and-decision-making-in-meetings/**.

ARTIFICIAL INTELLIGENCE SUPPORTS BRAINSTORMING In Chapter 12, we will introduce the use of bots. Some software allows users to create and post a bot (or avatar) that represents people in order to communicate anonymously. Artificial intelligence (AI) can also be used for pattern recognition and identifying ideas that are similar to each other. AI is also used in crowdsourcing (Section 11.7), which is used extensively for idea generation and voting.

Group Support Systems

A **group support system (GSS)**, which was discussed earlier, is any combination of hardware and software that enhances group work. GSS is a generic term that includes all forms of communication and collaborative computing. It evolved after information technology researchers recognized that technology could be developed to support many activities that normally occur at face-to-face meetings when they occur in virtual meetings (e.g., idea generation, consensus building, anonymous ranking). Also, a focus was made on collaboration rather than on minimizing conflicts.

A complete GSS is considered a specially designed information system software, but today, its special capabilities have been embedded in standard IT productivity tools. For example, Microsoft Office 365 includes Microsoft Teams (opening vignette). It also includes the tools for Web conferences. Also, many commercial products have been developed to support only one or two aspects of teamwork (e.g., videoconferencing, idea generation, screen sharing, wikis).

HOW GSS IMPROVES GROUP WORK The goal of GSS is to provide support to participants in improving the productivity and effectiveness of meetings by streamlining and speeding up the decision-making process and/or by improving the quality of the results. GSS attempts to increase process and task gains and decrease process and task losses. Overall, GSS has been successful in doing just that. Improvement is achieved by providing support to group members for the generation and exchange of ideas, opinions, and preferences. Specific features such as the ability of participants in a group to work simultaneously on a task (e.g., idea generation or voting) and anonymity produce improvements. The following are some specific GSS support activities:

- Supporting parallel processing of information and idea generation (brainstorming).
- Enabling the participation of larger groups with more complete information, knowledge, and skills.
- Permitting the group to use structured or unstructured techniques and methods.
- Offering rapid, easy access to external information.
- Allowing parallel computer discussions.
- Helping participants frame the big picture.
- Providing anonymity, which allows shy people to contribute to the meeting (i.e., to get up and do what needs to be done).
- Providing measures that help prevent aggressive individuals from controlling a meeting.
- Providing multiple ways to participate in instant anonymous voting.
- Providing structure for the planning process to keep the group on track.
- Enabling several users to interact simultaneously (i.e., conferencing).
- Recording all information presented at a meeting (i.e., providing *organizational memory*).

For GSS success stories, look for sample cases at vendors' Web sites. As you will see in many of these cases, collaborative computing led to dramatic process improvements and cost savings.

Note that only some of these capabilities are provided in a single package from one vendor.

SECTION 11.5 REVIEW QUESTIONS

1. Define *GDSS* and list the limitations of the initial GSS software.
2. List the benefits of GDSS.
3. List process gains made by GDSS.
4. Define *decision room*.
5. Describe Web-based GSS.
6. Describe how GDSS supports brainstorming and idea generation.

11.6 COLLECTIVE INTELLIGENCE AND COLLABORATIVE INTELLIGENCE

Groups or teams are created for several purposes. Our book concentrates on support for decision making. This section deals with the collective intelligence and collaborative intelligence of groups.

Definitions and Benefits

Collective intelligence (CI) refers to the total intelligence of a group. It is also refers to as the *wisdom of the crowd*. People in a group are using their skills and knowledge for solving problems and providing new insights and ideas. The major benefits are the ability to solve complex problems and/or design new products and services that result from innovations. A major research center on collective intelligence (CI) is the MIT Center for Collective Intelligence (CCI) (**cci.mit.edu**). A major study aspect of CCI is how people and computers can work together so that teams can be more innovative than any individual, group, or computer can be alone. CI appears in several disciplines ranging from sociology to political science. Our interest here is in CI as it relates to computerized decision making. We cover CI here and in Section 11.7 where we present the topic of crowdsourcing. In Section 11.8, we present swarm intelligence, which is also an application of CI. For the benefits of CI, see **50Minutes.com** (2017).

TYPES OF COLLECTIVE INTELLIGENCE One way to categorize CI is to divide it into three major areas of applications: *cognition, cooperation*, and *coordination*. Each of these can be further divided. For an overview, see collective intelligence on Wikipedia. Our interest is in applications by which the group synergy helps in problem solving and decision making. People contribute their experience and knowledge, and the group interactions and the computerized support help in making better decisions.

Thomas W. Malone, the founder and director of CCI at MIT, considers CI as a broad umbrella. He views collective intelligence as "groups of individuals acting collectively in ways that seem intelligent." The CCI work, known as the *Edge*, is available at the Edge video (31:45 minutes) available at **edge.org/conversation/thomas_w__malone-collective-intelligence**.

Computerized Support to Collective Intelligence

Collective intelligence can be supported by many of the tools and platforms described in Sections 11.4 and 11.5. In addition, the Internet, intranet, and the IoT (Chapter 13) play a major role in facilitating CI by enabling people to share knowledge and ideas.

Example 1: The Carnegie University Foundation Supports Network Collaboration

The Carnegie Foundation was looking for ways to have people work together collaboratively in order to accelerate improvements and to share data and learning across its networks of people. The solution is an online workspace called the Carnegie Hub, which serves as an access point to resources and enables engagement in group work and collaboration.

The Hub uses several software products, some of which were described in Section 11.4, such as Google Drive, creating a collaborative workspace. The major aspects of the Carnegie Collection Intelligence project follow:

1. Content is shared in one place (the "cloud") for everyone to view, edit, or contribute even at the same time.
2. All data and knowledge are stored in one location on the Web. Discovery is easy.
3. Asynchronous conversations using discussion boards are easy; all notes are publicly displayed, documented, and stored.
4. These aspects facilitate social collaboration, commitment to problem solving, and peer support. The Carnegie University faculty is now a community of practice, using collective intelligence to plan, create, and solve problems together. For details, see Thorn and Huang (2014).

Example 2: How Governments Tap IoT for Collective Intelligence

According to Bridgwater (2018), governments are using IoT to support decision making and policy creation. Governments are trying to collect information and knowledge from people and increasingly do so via IoT. Bridgwater cites the government of the United Arab Emirates that uses IoT to enhance public decision making. The IoT systems collect ideas and aspirations of the citizens. The collective intelligence platform allows the targeting of narrowly defined groups. Real estate plans are subjected to the opinion of residents in the vicinity of proposed developments. The country's project of smart cities is combined with CI (Chapter 13). In addition to IoT, there are activities in CI and networks as shown in Application Case 11.1.

Application Case 11.1 Collaborative Modeling for Optimal Water Management: The Oregon State University Project

Introduction

Water management is one of the most important challenges for many communities. In general, the demand for water is growing while the supply could shrink (e.g., due to pollution). Managing water requires the involvement of numerous stakeholders ranging from consumers and suppliers to local governments and sanitation experts. The stakeholders must work together. The objective is to have responsible water use and water preservation. The accounting office of PwC published report 150CO47, "Collaboration: Preserving Water Through Partnership That Works" available at **pwc.com/hu/hu/kiadvanyok/assets/pdf/pwc_water_collaboration.pdf.** It describes the problem and its benefits and risks. The report shares the different stakeholders' perspectives, identifies the success factors of collaboration, and weighs the trade-offs for evaluating alternative solutions for the water management issue. An interesting framework for a solution is the collaborative modeling developed at Oregon State University in collaboration with Indiana University-Purdue University.

The Challenge

Planning and managing water conservation activities are not simple tasks. The idea is to develop a user-friendly tool that will enable all stakeholders to participate in these activities. It is necessary to involve the stakeholder communities in using scientifically developed guidelines for designing water conservation practices. Here are some of the requirements of the desired tool:

- The tool needs to be interactive and human guided and operated.
- It needs to be Web-based and user friendly.
- Both individuals and groups should be able to use it.
- It should enable users to view and evaluate solution designs based on both quantitative and qualitative criteria.

The Solution: WRESTORE

Watershed Restoration Using Spatio-Temporal Optimization (WRESTORE) is a Web-based tool that meets the preceding requirements. It is based on AI and analytical optimization algorithms. The algorithms process dynamic simulation models and allow users to spatially optimize the location of new water conservations. In addition to using the dynamic simulation models, users are able to include their own personal subjective views and qualitative criteria. WRESTORE generates alternative practices that users can discuss and evaluate.

Incorporation of human preferences to computer solutions makes the solutions more acceptable. The AI part of the project includes machine learning and crowdsourcing (Section 11.7) to solicit information from the crowd. The reason for the participative collaboration is that water is an essential resource and should not be only centrally controlled. The AI technologies "democratize" water management while harnessing the power of people and computers to solve difficult water management problems.

The machine-learning algorithms learn from what people are doing. Human feedback helps AI to identify best solutions and strategies. Thus, humans and machines are combined to solve problems together.

The Results

WRESTORE developers are experimenting with the technology in several places and so far have achieved full collaboration from participating stakeholders. Initial results indicate the creation by WRESTORE of innovative ideas for developing water resources and distribution methods that save significant amounts of water.

Questions for Case 11.1

1. Crowdsourcing is used to find information from a crowd. Why is it needed in this case? (see Section 11.7 if you are not familiar with crowdsourcing).
2. How does WRESTORE act as a CI tool?
3. Debate centralized control versus participative collaboration. Cite the pros and cons of each.
4. Why it is difficult to manage water resources?
5. How can an optimization/simulation/AI model support group work in this case?

Sources: Compiled from Basco-Carrera et al. (2017), KTVZ.com (Channel 21, Oregon, March 21, 2018), and Babbar-Sebens et al. (2015).

How Collective Intelligence May Change Work and Life

For several decades, researchers studied the relationship of CI and work. For example, Doug Engebert, a pioneer in CI, describes how people work together in response to a shared challenge and how they can leverage their collective memory, perception, planning, reasoning, and so on into powerful knowledge. Since Engebert's pioneering work, the impact of technology is increasing organizations' CI and building *collaborative* communities of knowledge. In summary, CI attempts to augment human intelligence to solve business and social problems. This basically means that CI allows more people to have more engagement and involvement in organizational decision making. At MIT's CCI, research is done on how people and computers can work together to improve work (see also Section 11.9). MIT's CCI focuses on the role of networks, including the Internet, intranets, and IoT. Researchers there found that organizations' structures tend to be flatter, and more

decisions are delegated to teams. All this results in decentralized workplaces. For further discussion on MIT's CCI, see MIT's blog of April 3, 2016, at **executive.mit.edu/blog/will-collective-intelligence-change-the-way-we-work/**. For a comprehensive view on how CI can change the entire world, see Mulgan (2017).

A major thrust in CI is the collaboration efforts within a group, as described next.

Collaborative Intelligence

Placing people in groups and expecting them to collaborate with the help of technology may be wishful thinking. Management and behavioral researchers study the issue of how to make people collaborate in groups.

Called by some *collaborative intelligence*, Coleman (2011) stipulates that group collaboration has the following 10 components: (1) willingness to share, (2) knowing how to share, (3) being willing to collaborate, (4) knowing what to share, (5) knowing how to build trust, (6) understanding team dynamics, (7) using correct hubs for networking, (8) mentoring and coaching properly, (9) being open to new ideas, and (10) using computerized tools and technology. A similar list is provided at **thebalancecareers.com/collaboration-skills-with-examples-2059686**.

Computerized tools and technologies are critical enablers of communication, collaboration, and people's understanding of each other.

How to Create Business Value from Collaboration: The IBM Study

Groups and team members provide ideas and insights. To excel, organizations must utilize people's knowledge, some of which is created by collective intelligence. One way to do this is provided by a study of collective intelligence conducted by the IBM Institute for Business Value. The study is available (free) at **www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-collective-intelligence.html**. There is also a free executive summary. The study presents three major points:

1. CI can enhance organizational outcomes by correctly tapping the knowledge and experience of working groups (including customers, partners, and employees).
2. It is crucial to target and motivate the appropriate participants.
3. CI needs to address the issue of participants' resistance to change. All in all, IBM concludes that "Collective intelligence is a powerful resource for creating value using the experience and insights of vast numbers of people around the world."

   Access the untapped knowledge of your networks, IBM. **(www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-collective-intelligence.html)**

An offshoot of CI is crowdsourcing, the topic of the next section (11.7).

SECTION 11.6 REVIEW QUESTIONS

1. What is collective intelligence (CI)?
2. List the major benefits of CI.
3. How is CI supported by computers?
4. How can CI change work and life?
5. How can CI impact organization structure and decision making?
6. The Carnegie case described how standard collaboration tools create a collective intelligence infrastructure. The WRESTORE case described a modeling analytical framework that enables stakeholders to collaborate. What are the similarities and differences between the two cases?
7. Describe collaborative intelligence.
8. How do you create business value from collective intelligence?

11.7 CROWDSOURCING AS A METHOD FOR DECISION SUPPORT

Crowdsourcing refers to outsourcing tasks to a large group of people (crowd). One of the major reasons for doing so is the potential for the wisdom of a crowd to improve decision making and assist in solving difficult problems; see Power (2014). Therefore, crowdsourcing can be viewed as a method of *collective intelligence*. This section is divided into three parts: The essentials of crowdsourcing, crowdsourcing as a decision support mechanism, and implementing crowdsourcing for problem solving.

The Essentials of Crowdsourcing

Crowdsourcing has several definitions because it is used for several purposes in a number of fields. For a tutorial on crowdsourcing and examples, view the video (14:51 min.) at **youtube.com/watch?v=lXhydxSSNOY**. Crowdsourcing means that an organization is outsourcing or farming out work for several reasons: Necessary skills may not be available internally, speed of execution is needed, problems are too complex to solve, or special innovation is needed.

SOME EXAMPLES

- Since 2005, Doritos Inc. has run a "Crash the Super Bowl" contest for creating a 30-second video for the Super Bowl. The company has given $7 million in prizes in the last 10 years for commercials composed by the public.
- Airbnb is using user-submitted videos (15 seconds each) that describe travel sites.
- Dell's Idea Storm (**ideastorm.com**) enables customers to vote on features of Idea Storm the customers prefer, including new ones. Dell is using a technically oriented crowd, such as the Linux (**linux.org**) community. The crowd submits ideas and sometimes members of the community vote on them.
- Procter & Gamble's researchers post their problems at **innocentive.com** and **ninesigma.com**, offering cash rewards to problem solvers. It uses other crowdsourcing service providers such as **yourencore.com**.
- The LEGO company has a platform called LEGO Ideas through which users can submit ideas for new LEGO sets and vote on submitted ideas by the crowd. Accepted ideas generate royalties to those who proposed them if the ideas are commercialized.
- PepsiCo solicits ideas regarding new potato chip flavors for the company's Lay's brand. Over the years, the company has received over 14 million suggestions. The estimated contribution to sales increase is 8 percent.
- Cities in Canada are creating real-time electronic city maps to inform cyclists about high-risk areas to make the streets safer. Users can mark the maps when they experience a collision, bike theft, road hazard, and so on. For details, see Keith (2018).
- U.S. intelligence agencies have been using ordinary people (crowds) to predict world events ranging from the results of elections to the direction of prices.
- Hershey crowdsourced potential solutions of how to ship chocolate in warm climates. For how this was done, see Dignan (2016). The winning prize was $25,000.

These examples illustrate some of the benefits of crowdsourcing, such as wide exposure to expertise, increased performance and speed, and improved problem-solving and innovation capabilities. These examples also illustrate the variety of applications.

MAJOR TYPES OF CROWDSOURCING Howe (2008), a crowdsourcing pioneer, divided the crowdsourcing applications into the following types (or models):

1. **Collective intelligence (or wisdom).** People in crowds are solving problems and providing new insights and ideas leading to product, process, or service innovations.
2. **Crowd creation.** People are creating various types of content and sharing it with others (for pay or free). The created content may be used for problem

solving, advertising, or knowledge accumulation. Content creation can also be done by splitting large tasks into small segments (e.g., contributing content to create Wikipedia).
3. **Crowd voting.** People are giving their opinions and ratings on ideas, products, or services, as well as evaluating and filtering information presented to them. An example is voting in *American Idol* competitions.
4. **Crowd support and funding.** People are contributing and supporting endeavors for social or business causes, such as offering donations, and micro-financing new ventures.

Another way to classify crowdsourcing is by the type of work it does. Some examples with a crowdsourcing vendor for each follow:

- Logo design—Design Bill
- Problem solving—InnoCentive, NineSigma, IdeaConnection
- Business innovation—Chardix
- Brand names—Name This
- Product and manufacturing design—Pronto ERP
- Data cleansing—Amazon Mechanical Turk
- Software testing—uTest
- Trend watching—TrendWatching
- Images—Flickr Creative Commons

For a compressive list of crowdsourcing, collective intelligence, and related companies, see **boardofinnovation.com**.

THE PROCESS OF CROWDSOURCING The process of crowdsourcing differs from application to application, depending on the nature of the specific problem to be solved and the method used. However, the following steps exist in most enterprise crowdsourcing applications, even though the details of the execution may differ. The process is illustrated in Figure 11.3.

1. Identify the problem and the task(s) to be outsourced.
2. Select the target crowd (if not an open call).
3. Broadcast the task to the crowd (or to an unidentified crowd in an open call).
4. Engage the crowd in accomplishing the task (e.g., idea generation, problem solving).
5. Collect user-generated content.
6. Have the quality of submitted material evaluated by the management that initiated the request, by experts, or by a crowd.
7. Select the best solution (or a short list).
8. Compensate the crowd (e.g., the winning proposal).
9. Implement the solution.

Note that we show the process as sequential, but there could be loops returning to previous steps.

Crowdsourcing for Problem-Solving and Decision Support

Although there are many potential activities in crowdsourcing, major ones are supporting the managerial decision-making process and/or providing a solution to a problem. A complicated problem that is difficult for one decision maker or a small group to solve may be solved by a crowd, which can generate a large number of ideas for solving a

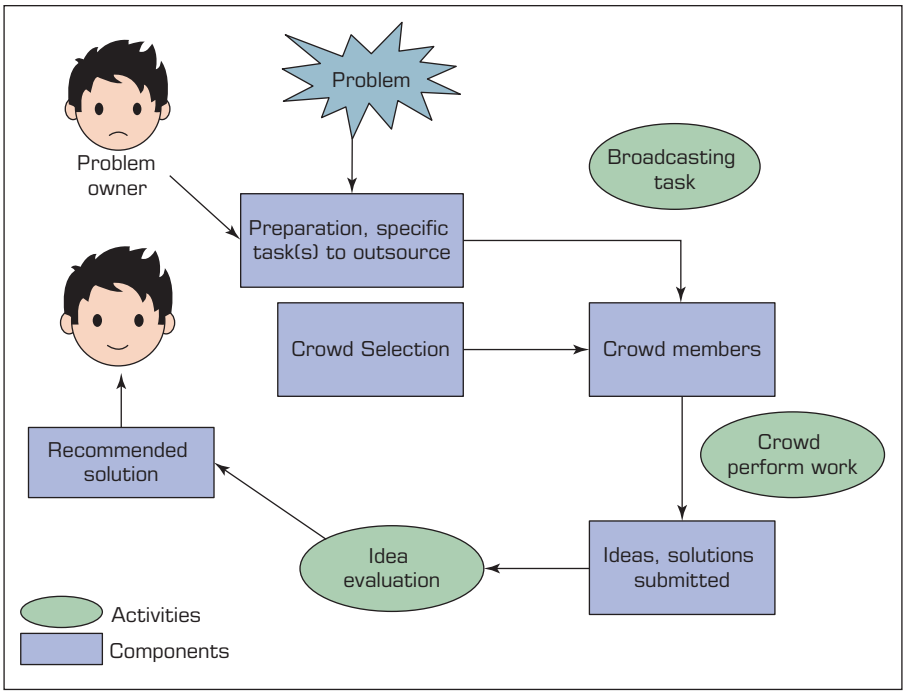


FIGURE 11.3 **The Crowdsourcing Process.**

problem. However, inappropriate use of crowdsourcing could generate negative results (e.g., see Grant, 2015). On how to avoid the potential pitfalls of crowdsourcing, see Bhandari et al., 2018.

THE ROLE OF CROWDSOURCING IN DECISION MAKING Crowds can provide ideas in a collaborative or a competitive mode. However, the crowd's role may differ at different stages of the decision-making process. We may use a crowd to decide how to respond to a competitor's act or to help us decide whether a proposed design is useful. Chiu et al. (2014) adopted Herbert Simon's decision-making process model to outline the potential roles of a crowd. Simon's model includes three major phases before implementation: *intelligence* (information gathering and sharing for the purpose of problem solving or opportunity exploitation, problem identification, and determination of the problem's importance), *design* (generating ideas and alternative solutions), and *choice* (evaluating the generated alternatives and then recommending or selecting the best course of action). Crowdsourcing can provide different types of support to this managerial decision-making process. Most of the applications are in the design phase (e.g., idea generation and co-creation) and in the choice phase (voting). In some cases, support can be provided in all phases of the process.

Implementing Crowdsourcing for Problem Solving

While using an open call to the public can be done fairly easily by the problem owner, people who need to solve difficult problems usually like to reach experts for solving problems (solvers). For a company to obtain assistance in finding such experts, especially externally, it can use a third-party vendor. Such vendors have hundreds of thousands or even millions of preregistered solvers. Then, the vendor can do the job as illustrated in Application Case 11.2.

Application Case 11.2 How InnoCentive Helped GSK Solve a Difficult Problem

GlaxoSmithKline (GSK) is a UK-based global pharmaceutical/healthcare company, with over 100,000 employees. The company strives on innovations. However, despite its mega size and global presence, it has problems that it needs outside expertise to solve.

The Problem

The company researched a potentially disruptive technology that promised cure to difficult diseases. The company wanted to discover which disease to use as a test bed for the potential innovative treatments. It was necessary to make sure that the selection will cover a disease where every aspect of the new treatment is checked. Despite its large size, GSK wanted some outside expertise to support and check the in-house research efforts.

The Solution

GSK decided to crowdsource the problem solution to experts, using InnoCentive Corp. (**Innocentive.com**). InnoCentive is a US-based global crowdsourcing company. The company receives challenges from client companies like GSK. These challenges are posted for solvers to see with the potential rewards, in InnoCentive's Challenge Center. Solvers that think they want to participate follow instructions and may sign an agreement. The solutions submitted are evaluated, and awards are provided to the winners.

The GSK Situation

In total, 397 solvers engaged in this challenge, even the reward was minimal ($5000). The solvers resided in several countries. The solvers submitted 66 proposed solutions. The entire process lasted 75 days.

The Results

The winning solution proposed a new area that was not considered by GSK teams. The proposer was a Bulgarian who based his idea on a Mexican publication. Several other winning proposals contributed useful ideas. Also, the process enabled collaboration between the GSK team and the winning researchers.

Questions for Case 11.2

1. Why did GSK decide to crowdsource?
2. Why did the company use InnoCentive?
3. Comment on the global nature of the case.
4. What lessons did you learn from this case?
5. Why do you think a small $5000 reward is sufficient?

Sources: Compiled from InnoCentive Inc. Case Study GlaxoSmithKline. Waltham, MA., GSK Corporate Information (**gsk.com**) and **InnoCentive.com/our-solvers/**.

CROWDSOURCING FOR MARKETING More than 1 million customers are registered at Crowd Tap, the company that provides a platform named Suzy that enables marketers to conduct crowdsourcing studies.

SECTION 11.7 REVIEW QUESTIONS

1. Define *crowdsourcing*.
2. Describe the crowdsourcing process.
3. List the major benefits of the technology.
4. List some areas for which crowdsourcing is suitable.
5. Why may you need a vendor to crowdsource the problem-solving process?

11.8 ARTIFICIAL INTELLIGENCE AND SWARM AI SUPPORT OF TEAM COLLABORATION AND GROUP DECISION MAKING

AI, as seen in Chapter 2, is a diversified field. Its technologies can be used to support group decision making and team collaboration.

AI Support of Group Decision Making

A major objective of AI is to automate decision making and/or to support its process. This objective holds also for decisions made by groups. However, we cannot automate a decision made by a group. All we can do is to support some of the steps in a group's decision-making process.

A logical place to start is Figure 11.1. We can examine the different steps of the process and see where AI can be used.

1. *Meeting preparation.* AI is used to find a convenient time for meetings to take place. AI can assist in scheduling meetings so that all can participate.
2. *Problem identification.* AI technologies are used for pattern recognition that can identify areas that need attention. AI can be used in other types of analysis to identify potential or difficult to pinpoint problems.
3. *Idea generation.* AI is known for its quest for creativity. Team members can increase their creativity when they use AI for support.
4. *Idea organization.* Natural language processing (NLP) can be used to sort ideas and organize them for improved evaluation.
5. *Group interaction and collaboration.* AI can facilitate communication and collaboration among group members. This activity is critical in the process of arriving at a consensus. Also, Swarm AI (see the end of this section) is designed to increase interactions among group members so their combined wisdom is elevated.
6. *Predictions.* AI supports predictions that are required to assess the impact of the ideas generated regarding performance and/or impacts in the future. Machine learning, deep learning, and Swarm AI are useful tools in this area.
7. *Multinational groups.* Collaboration among people located in different countries is on the rise. AI enables group interaction of people who speak different languages, in real time.
8. *Bots are useful in supporting meetings.* Group members may consult Alexa and other bots. Chatbots can provide answers to queries in real time.
9. *Other advisors.* IBM Watson can provide useful advice during meetings, supplementing knowledge provided by participants and by Alexa.

Example

In 2018, Amazon.com was looking for a site for its second headquarters. A robot named Aiera from Wells Fargo Securities used deep learning to predict that the winning site would be Boston (Yurieff, 2018a). (When this chapter was written, the decision had not been made.)

For an academic approach on how to improve group decision making by AI, see Xia (2017).

AI Support of Team Collaboration

Organizations today are looking for ways to increase and improve collaboration with employees, business partners, and customers. To gain insight into how AI may impact collaboration, Cisco Systems sponsored a global survey, AI Meets Collaboration (Morar HPI, 2017), regarding the impact of AI, including the use of virtual assistants in the work space. The major findings of this survey are:

1. Virtual assistants increase employees' productivity, creativity, and job satisfaction. Bots also enable employees to focus on high-value tasks.
2. Bots are accepted as part of workers' teams.

3. Bots improve conference calls. They also can take meetings notes and schedule meetings.
4. AI can use facial recognition to sign in eligible people to meetings.
5. Personal characteristics are likely to influence how people feel about AI in the workplace.
6. Employees in general like to have AI in their teams.
7. Security is a major concern when AI, such as virtual assistants, is used in teams.
8. The major AI tools that are most useful are NLP and voice response; AI can also summarize the key topics of meetings and understand participants' needs. AI can be aware of organizational goals and workers' skills and can make suggestions accordingly.

For how virtual meetings are supported with AI by Cisco Systems in their leading products, see Technology Insight 11.2.

TECHNOLOGY INSIGHT 11.2 How Cisco Improves Collaboration with AI

Cisco Systems is well known for its collaboration products such as Spark and Webex. The first step in introducing AI was to acquire MindMeld's AI platform for use in Cisco's collaboration products. The project's objective was to improve the conversational interferences for any application or device so users could better understand the context of conversations. MindMeld uses machine learning to improve the accuracy of voice and text communication. To do so, it uses NLP and five varieties of machine learning. Cisco is also integrating IBM Watson into its enterprise collaboration solutions. As you may recall from Chapter 6, Watson is a powerful advisor. AI collaboration tools can increase efficiency, speed idea generation, and improve the quality of decisions made by groups. The improved Cisco's technology will be used in conference rooms and everywhere else. One of the major AI projects is the assistant to Spark.

Monica, a Digital Assistant to the Spark Collaboration Platform

Monica is trained to answer users' queries by employing machine learning. Furthermore, users can use Monicait to interact with the Spark collaboration platform using natural language commands. It is an enterprise assistant similar to Alexa and Google Assistant (Chapter 12). Cisco's Monica is the world's first enterprise-ready voice assistant specifically designed to support meetings. The bot has deep-domain conversational AI that adds cognitive capabilities to the Spark platform.

Monica can assist users in several of the steps of Figure 11.1, such as:

- Organize meetings.
- Provide information to participants before and during meetings.
- Navigate and control Spark's devices.
- Help organizers find a meeting room and reserve it.
- Help share screens and bring up a whiteboard.
- Take meeting notes and organize them.

In the near future, Monica will know about participants' internal and external activities and will schedule meetings using this information. Additional functions to support more steps of the process in Figure 11.1 will be added in the future.

For more about the assistant, see **youtube.com/watch?v=8OcFSEbR_6k** (5:10 minutes).

Note: Cisco Spark will become Webex Teams with more AI functionalities. In addition, Webex meetings will include videoconferencing for collaboration and other supports to meetings.

Sources: Compiled from Goecke (2017), Finnegan (2018), and Goldstein (2017).

Swarm Intelligence and Swarm AI

The term **swarm intelligence** refers to the collective behavior of decentralized, self-organized systems, natural or artificial (per Wikipedia). Such systems consist of things (e.g., ants, people) interacting with each other and their environment. A swarm's actions are not centrally controlled, but they lead to intelligent behavior. In nature, there are many examples (e.g., ant colonies, fish schools) of such behaviors.

Natural groups were observed to amplify their group intelligence by forming swarms. Social creatures, including people, can improve the performance of their individual members when working together as a unified system. In contrast with animals and other species whose interactions among group members are natural, people need technology to exhibit swarm intelligence. This concept is used in studies and implementation of AI and robotics. The major applications are in the area of predictions.

Example

A study at Oxford University (United Kingdom) involved predicting the results of all 50 English Premier League soccer games over five weeks. A group of independent judges scored 55 percent accuracy when working alone. However, when predicting using an AI swarm, their prediction success increased to 72 percent (an improvement of 31 percent). Similar improvement was recorded in several other studies.

In addition to improved prediction accuracy, studies show that using swarm AI results in more ethical decisions than that of individuals (Reese, 2016).

SWARM AI TECHNOLOGY Swarm AI (or AI swarm) provides the algorithms for the interconnections among people creating the human swarm. These connections enable the knowledge, intuition, experience, and wisdom of individuals to merge into single improved swarm intelligence. Results of swarm intelligence can be seen in the TED presentation (15:58 min.) at **youtube.com/watch?v=Eu-RyZt_Uas**. Swarm AI is used by several third-party companies (e.g., Unanimous.aI, as illustrated in Application Case 11.3.

Application Case 11.3 XPRIZE Optimizes Visioneering

XPRIZE is a nonprofit organization that allocates prizes via competitions to promote innovations that have the potential to change the world for the better. The main channel for designing prizes that solve humanity's grandest challenges is called Visioneering. It attempts to harness the power of the global crowd to develop solutions to important challenges. The organization's major event is an annual summit meeting where prizes are designed and proposals are evaluated. The experts at XPRIZE develop concepts and turn them into incentivized competitions. Prizes are donated by leading corporations.

For example, in 2018, IBM Watson donated a $5 million prize called "AI approaches and collaboration." The competition had 142 registered teams, and 62 were left in round 2 in June 2018. The teams are invited to create their own goals and solutions to a grand challenge.

The Problem

Every year, there is a meeting of 250 members of "Visioneers Summit Ideation" where top experts (entrepreneurs, politicians, scientists, etc.), participate to discover and prioritize topics for the XPRIZE agenda.

Finding the top global problems can be a very complex challenge due to a large number of variables. In just a few days, top experts need to use their collective wisdom to agree on the next year's XPRIZE top challenges. The method used to support the group's decision is a critical success factor.

(*Continued*)

Application Case 11.3 (Continued)

The Solution

In the 2017 annual meeting for determining what challenge to use for 2018, the organization used the swarm AI platform (from Unanimous AI). Several small groups (swarms) moderated by AI algorithms were created to discover challenging topics. The mission was to explore ideas and agree on preferred solutions. The objective was to use the talents and brainpower of the participants.

In other words, the objective was to use the thinking together feature of swarm AI to generate each group's synergy with the AI algorithms acting as moderators. This way, smarter decisions were generated by the groups than its individual participants. The different groups examined six preselected topics: energy and infrastructure, learning human potential, space and new frontiers, plant and environment, civil society, and health and well-being. The groups brainstormed the issues. Then, each participant created a customized evaluation table. The tables were combined and analyzed by algorithms.

The Swarm AI replaced traditional voting methods by optimizing the detailed contribution of each participant.

The Results

Use of swarm AI did the following:

- Supported the generation of optimized answers and enabled fast buy-in from the participants.
- Enabled all participants to contribute.
- Provided a better voting system than in previous years.

Questions for Case 11.3

1. Why is the group discussion in this case complex?
2. Why is getting a consensus when top experts are involved more difficult than when non-experts are involved?
3. What was the contribution of swarm AI?
4. Compare simple voting to swarm AI voting.

Sources: Compiled from Unanimous AI (2018), **xprize.org**, and **xprize.org/about**.

SWARM AI FOR PREDICTIONS Swarm AI was used by Unanimous AI for making predictions in difficult-to-assess situations. Examples are:

- Predicting Super Bowl #52 number of points scored (used for spread waging).
- Predicting winners in the regular NFL season.
- Predicting the top four finishers of the 2017 Kentucky Derby.
- Predicting the top recipients of the Oscars in 2018.

SECTION 11.8 REVIEW QUESTIONS

1. Relate the use of AI to the activities in Figure 11.1.
2. Discuss the different ways that AI can facilitate group collaboration.
3. How can AI support group evaluation of ideas?
4. How can AI facilitate idea generation?
5. What is the analogy of swarm AI to swarms of living species?
6. How is swarm AI used to improve group work and to initiate group predictions?

11.9 HUMAN–MACHINE COLLABORATION AND TEAMS OF ROBOTS

Since the beginning of the Industrial Revolution, people and machines have worked together. Until the late 1900s, the collaboration was in manufacturing. But since then, due to advanced technology and changes in the nature of work, human–machine collaboration has spread to many other areas, including performing mental and cognitive work

and collaborating on managerial and executive work. According to Nizri (2017), human and AI collaboration will shape the future of work (see also Chapter 14).

Humans and machines can collaborate in many ways, depending on the tasks they perform. The collaboration with robots in the manufacturing scenario is an extension of the older model in which humans and robots collaborated with humans controlling and monitoring production and robots doing physical work that requires speed, power, accuracy, or nonstop attention. Robots are also doing work in hazardous environments. In general, robots complement human capabilities. An example is Amazon's distribution centers where over 50,000 mobile robots do a variety of tasks, mostly in hauling materials and helping to fulfill customer orders. The robotic technology enables fully collaborative solutions. For details, watch the video at Kuka **kuka.com/en-us/technologies/human-robot-collaboration**. Kuka's system allows the execution of complex jobs that can be done cost effectively.

Another collaborative human-robotic system is called YuMi. To see this system (from ABB Robotics) at work, watch the 4:38 min. video at **youtube.com/watch?v=2KfXY2SvlmQ**. Notice that the robot has two arms.

Human–Machine Collaboration in Cognitive Jobs

Advancement in AI enables the automation of nonmanual activities. While some intelligent systems are fully automated (see automated decision making in Chapter 2 and chatbots in Chapter 12), there are many more examples of human–machine collaboration in cognitive jobs (e.g., in marketing and finance). An example is in investment decisions. A human asks the computer for advice concerning investments, and after receiving the advice, can ask more questions, changing some of the input. The difference from the past is that today the computers (machines) can provide much more accurate suggestions, by using machine learning and deep learning. Another collaboration example involves medical diagnoses of complex situations. For example, IBM Watson provides medical advice, which permits doctors and nurses to significantly improve their jobs. Actually, the entire field of machines advising humans is reaching new heights. For more on the increasing collaborative power of AI, see Carter (2017).

TOP MANAGEMENT JOBS A major task of managers is decision making, which has become one area of human–machine collaboration. Use of AI and analytics has improved decision making considerably, as illustrated throughout this book. For an overview, see Wladawsky-Berger (2017).

McKinsey & Company and MIT are two major players in researching the topic of collaboration between managers and machines. For example, Dewhurst and Wilmott (2014) report on its increased use of man-machine collaboration, using deep learning. A Hong Kong company even appointed a decision-making algorithm to its board of directors. Companies are using crowdsourcing advice to support complex problem solving, as illustrated in Section 11.7.

Robots as Coworkers: Opportunities and Challenges

Sometime in the future, walking and talking humanoid robots will socialize with humans during breaks from work. Someday, robots will become cognitive coworkers and help people be more productive (as long as people do not talk too much with the robots).

According to Tobe (2015), a study at a BMW factory found that human–robot collaboration could be more productive than either humans or robots working by themselves. Also, the study found that collaboration reduced idle time by 85 percent. This is because people and machines capitalize on the strengths of each (Marr, 2017).

The following challenges must be considered:

- Designing a human–machine team that capitalizes on the strength of each partner.
- Exchanging information between humans and robots.
- Preparing company employees in all departments for the collaboration (Marr, 2017).
- Changing business processes to accommodate human–robot collaboration (Moran, 2018).
- Ensuring the safety of robots and employees that work together.

TECHNOLOGIES THAT SUPPORT ROBOTS AS COWORKERS Yurieff (2018b) lists the following examples of facilitating or considering robots as coworkers.

1. Virtual reality can be used as a powerful training tool (e.g., for safety).
2. A robot is working with an ad agency in Japan to generate ideas.
3. A robot can be your boss.
4. Robots are coworkers in providing parts out of bins in assembly lines and can check quality together with humans.
5. AI tools measure blood flow and volume of the cardiac muscles in seconds (instead of minutes when done completely by a radiologist). This information facilitates the decisions made by radiologists.

BLENDING HUMANS AND AI TO BEST SERVE CUSTOMERS Genesys Corp. commissioned Forrester Research Company to conduct a global study in 2017 to find how companies are using AI to improve customer service. The study, titled "Artificial Intelligence with the Human Touch," is available at no charge from **genesys.com/resources/artificial-intelligence-with-the-human-touch**. A related video is available at **youtube.com/watch?v=NP2qqwGTNPk**.

The study revealed the following:

1. "AI is already transforming enterprises by increasing worker efficiency and productivity, delivering better customer experiences and uncovering new revenue streams" (from the Executive Summary).
2. A major objective of man–machine collaboration is to improve the satisfaction of both customers and companies' agents rather than reduce cost.
3. Human agents' ability to connect emotionally with customers for the increased satisfaction of themselves and customers is superior to that of service provided by AI.
4. By blending the strengths of humans and AI, companies achieve better customer service satisfaction of customers (71 percent) and agents (69 percent).

Note that AI excels in the support of marketing and advertising as illustrated in Chapter 2. See also Loten (2018) for the use of AI to support customer relationship management (CRM) and of crowdsourcing and collective intelligence to support marketing.

COLLABORATIVE ROBOTS (CO-BOTS) Collaborative robots (co-bots) are designed to work with people, assisting in executing various tasks. These robots are not very smart, but their low cost and high usability make them popular. For details, see Tobe (2015).

Teams of collaborating Robots

One of the future directions in robotics is creating teams of robots that are designed to do complex work. Robot teams are common in manufacturing where they serve each other or join a robot group in simple assembly jobs. An interesting example is the use of a team of robots in preparation to land on Mars.

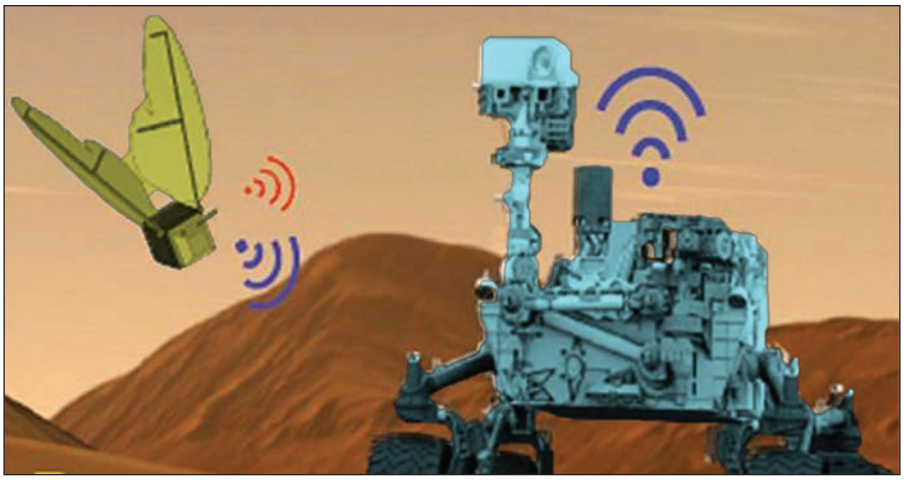


FIGURE 11.4 **Team of Robots Prepares to Go to Mars.** *Source:* C. Kang.

Example: Teams of Robots to Explore Mars

Before people land on Mars, scientists need to know more about the "Red Planet." The idea was to use teams of robots. The German Research Centers for Artificial Intelligence (DFKI) conducted simulation experiments in the desert of Utah. The details of this simulation are described by Staff Writers (2016). The process is illustrated in a 4:54 min. video at **youtube.com/watch?v=pvKIzldni68/** showing robots' collaboration. For more information, see **robotik.dfki-bremen.de/en/research/projects/ft-utah.html**.

DFKI is not the only entity that plans to explore the surface of Mars. NASA plans to send swarms of robot bees with flapping wings called *Marsbees* that will operate in a group to explore the land and air of the Red Planet. The reason for the flapping wings structure is to enable low-energy flights (like bumblebees). Each robot is the size of a bee. Part of a wireless communication network, Marsbees will together create networks of sensors. Information will be delivered to a mobile base (see Figure 11.4, showing one robot) that will be the main communication center and a recharging station for the Marsbees. For more information, see Kang (2018).

Getting robots to work together is being researched at MIT. They use their perception system to sense the environment, and then they communicate their findings to each other and coordinate their work. For example, a robot can open a door for another robot. Read about how this is done and watch a video at **ft.com/video/ea2d4877-f3fb-403d-84a8-a4d2d4018c5e**.

Example

Alibaba.com is using teams of robots in its smart warehouses where robots do 70 percent of the work. This is shown in a video at **youtube.com/watch?v=FBl4Y55V2Z4**.

Social collaboration of robots is being investigated by watching the behavior of swarms of ants and other species to learn how to design robots to work in teams. Watch the TED presentation at **youtube.com/watch?v=ULKyXnQ9xWA** on how to design a robot collaboration.

Having robots collaborate involves several issues such as making sure they do not collide with each other. This is a part of the safety issue regarding robotics. Finally, you can build your own team of robots with LEGO's Mindstorms. For details, see Hughes and Hughes (2013).

SECTION 11.9 REVIEW QUESTIONS

1. Why is there an increase in human–machine collaboration?
2. List some benefits of such collaboration.
3. Describe how collaborating robotics can be used in manufacturing.
4. Discuss the use of teams of robots.
5. What will do robots on Mars?

Chapter Highlights

- *Groupware* refers to software products that provide collaborative support to groups (including conducting meetings).
- Groupware can support decision-making and problem solving directly or indirectly by improving communication between team members.
- People collaborate in their work (called *group work*). Groupware (i.e., collaborative computing software) supports group work.
- Group members may be in the same organization or in different organizations in the same or in different locations and may work at the same or different times.
- The time/place framework is a convenient way to describe the communication and collaboration patterns and support of group work. Different technologies can support different time/place settings.
- Working in groups can result in many benefits, including improved decision making, increased productivity and speed, and cost reductions.
- Communication can be synchronous (i.e., same time) or asynchronous (i.e., sent and received at different times).
- The Internet, intranets, and IoT support virtual meetings and decision making through collaborative tools and access to data analysis, information, and knowledge.
- Groupware for direct support typically contains capabilities for brainstorming, conferencing, scheduling group meetings; planning; resolving conflicts; videoconferencing; sharing electronic documents; voting; formulating policy; and analyzing enterprise data.
- A GDSS is any combination of hardware and software that facilitates decision-making meetings. It provides direct support in face-to-face settings and in virtual meetings, attempting to increase process gains, and reducing process losses of group works.
- Collective intelligence is based on the premise that the combined wisdom of several collaborating people is greater than that of individuals working separately.
- Each of the several configurations of collective intelligence can be supported differently by technology.
- Several collaboration platforms, such as Microsoft Teams and Slack, can facilitate collective intelligence.
- Idea generation and brainstorming are key activities in group work for decision making. Several collaboration software and AI programs are supporting these activities.
- Crowdsourcing is a process of outsourcing work to a crowd. Doing so can improve problem solving, idea generation, and other innovative activities.
- Crowdsourcing can be used to make predictions by groups of people, including crowds. Results have shown better predictions, especially when communication is used among the predictors than when no communication was enabled.
- One method of communication in crowdsourcing is based on swarm intelligence. A technology known as *swarm AI* has had significant success.
- AI can support many activities in group decision making.
- Human–machine collaboration can be a major method of work in the future.
- Machines that once supported manufacturing work are used now also in support of cognitive, including managerial, work.
- For people and machines to work in teams, it is necessary to make special preparations.
- Robots may work in exclusive teams. They do so in manufacturing and possibly in other activities (e.g., explore Mars) as they become more intelligent.

Key Terms

asynchronous
brainstorming
collective intelligence
collaborative workspace
crowdsourcing
decision room
group decision making
group decision support system (GDSS)
group support system (GSS)
groupthink
groupware
group work
idea generation
online workspace
process gain
process loss
swarm intelligence
synchronous (real-time)
virtual meeting
virtual team

Questions for Discussion

1. Explain why it is useful to describe group work in terms of the time/place framework.
2. Describe the kinds of support that groupware can provide to decision makers.
3. Explain why most groupware is deployed today over the Web.
4. Explain in what ways physical meetings can be inefficient. Explain how technology can make meetings more effective.
5. Explain how GDSS can increase some benefits of collaboration and decision making in groups and eliminate or reduce some losses.
6. The initial term for group support system (GSS) was group decision support system (GDSS). Why was the word *decision* dropped? Does this make sense? Why, or why not?
7. Discuss why Microsoft SharePoint is considered a workspace. What kind of collaboration does it support?
8. Reese (2017) claims that swarm AI can be used instead of polls for market research. Discuss the advantages of swarm AI. In what circumstances would you prefer each method? (Read "Polls vs. Swarms" at Unanimous AI.)
9. What is a collaborative robot? What is an uncollaborative one?
10. Discuss the ways in which social collaboration can improve work in a digital workplace.
11. Provide an example of using analytics to improve decision making in sport.

Exercises

1. Go to **realtimeboard.com**. How can the site support idea creation and brainstorming?
2. Investigate how researchers are trying to develop collaborative computer systems that portray or display nonverbal communication factors (e.g., images).
3. For each of the software packages Skype Business and WebEx, check the trade literature and the Web for details and explain how each includes computerized collaborative support system capabilities.
4. Compare Simon's four-phase decision-making model to the steps in using GDSS.
5. A major claim in favor of wikis is that they can replace e-mail, eliminating its disadvantages (e.g., spam). Go to **socialtext.com** and review such claims. Find other supporters of switching to wikis. Then find counter arguments and conduct a debate on the topic.
6. Search the Internet to identify sites that describe methods for making meetings more effective and efficient.
7. Enter MIT Center for CI and review some of its recent activities. Write a report.
8. Debate the issue of the quality of crowdsourcing results. Start by viewing **youtube.com/watch?v=JJHAHQmiI3c**.
9. Find information about Yammer (a Microsoft company). Why is it considered a social collaboration tool? Why is it popular? Write a report.
10. Enter Dropbox.com and find its collaboration tools. Write a summary.
11. Read Pena (2017). Examine the 12 benefits of collaboration. Which are related to social collaboration?
12. Compare Microsoft's Universal Translator to Google's Translator. Concentrate on face-to-face conversation in real time.
13. Write a report on the issue of whether crowdsourcing produces superior decisions. Use Quora for help. Find other sources.
14. Investigate the status of IBM Connections Cloud. Examine all the collaboration and communication features. How does the product improve productivity? Write a report.

15. Compare Microsoft Teams to Spark Teams. Write a report.
16. Enter **crowdtap.com** and read Kurzer (2018) paper. Explain how the platforms work. Relate the material about crowdsourcing and collective intelligence. Write a report.
17. Go to **technologyreview.com** and look at the May 8, 2017, video (17:42 min.) "Next Generation Human-Machine Collaboration." Write a report.

References

Babbar-Sebens, M., et al. "A Web-Based Software Tool for Participatory Optimization of Conservation Practices in Watersheds." *Environmental Modelling & Software*, 69, 111–127, July 2015.

Basco-Carrera, L., et al. "Collaborative Modelling for Informed Decision Making and Inclusive Water Development." *Water Resources Management*, 31:9, July 2017.

Bhandari, R., et al. "How to Avoid the Pitfalls of IT Crowdsourcing to Boost Speed, Find Talent, and Reduce Costs." *McKinsey & Company*, June 2018.

Bridgwater, A. "Governments to Tap IoT for 'Collective Intelligence.'" *Internet of Business*, January 2, 2018.

Carter, R. "The Growing Power of Artificial Intelligence in Workplace Collaboration." *UC Today*, June 28, 2017.

Chiu, C-M., T. P. Liang, and E. Turban. "What Can Crowdsourcing Do for Decision Support?" *Decision Support Systems*, September 2014.

Coleman, D. "10 Components of Collaborative Intelligence." *CMS Wire*, November 21, 2011.

de Lares Norris, M. A. "Collaboration Technology Is the Driving Force for Productivity and Businesses Need to Embrace It . . . Now." *IT ProPortal*, January 4, 2018.

DeSanctis, G., and R. B. Gallupe. "A Foundation for the Study of Group Decision Support Systems." *Management Science,* 33:5, 1987.

Dewhurst, M., and P. Willmott. "Manager and Machine: The New Leadership Equation." *McKinsey & Company*, September 2014.

Dignan, L. "A Sweet Idea: Hershey Crowdsourcing for Summer Chocolate Shipping Concepts." *ZDNet*, January 14, 2016.

Digneo, C. "49 Online Collaboration Tools to Help Your Team Be More Productive." *Time Doctor*, 2018. **biz30.timedoctor.com/online-collaboration-tools/** (accessed July 2018).

50Minutes.com. *The Benefits of Collective Intelligence: Make the Most of Your Team's Skills*. Brussels, Belgium: **50Minutes.com** (Lemaitre Publishing), 2017.

Finnegan, M. "Cisco Shakes Up Collaboration Efforts; Morphs Spark into Webex." *Computer World*, May 2, 2018.

Goecke, J. "Meet Cisco Spark Assistant, Your Virtual Assistant for Meetings." *Cisco Blogs*, November 2, 2017.

Goldstein, P. "How Can AI Improve Collaboration Technology?" *Biztech Magazine*, June 5, 2017.

Grant, R. P. "Why Crowdsourcing Your Decision-Making Could Land You in Trouble." *The Guardian*, March 10, 2015.

Howe, J. *Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business*. New York: Crown Business, 2008.

Hughes, C., et al. *Build Your Own Teams of Robots with LEGO® Mindstorms® NXT and Bluetooth®*. New York, NY: McGraw-Hill/Tab Electronic, 2013.

Kang, C-K. "Marsbee—Swarm of Flapping Wing Flyers for Enhanced Mars Exploration." *NASA.gov*, March 30, 2018.

Keith, E. "Here's How a New Crowd-Sourced Map Is Making Canadian Streets Safer for Cyclists." *Narcity.com*, June 2018.

Kurzer, R. "Meet Suzy: The New Crowd Intelligence Platform with the Cute Name." *MarTech Today*, March 27, 2018.

Loten, A. "The Morning Download: AI-Enabled Sales Tools Spotlight Data Needs." *The Wall Street Journal*, March 27, 2018.

Marr, B. "Are You Ready to Meet Your Intelligent Robotic Co-Worker?" *Forbes.com*, September 8, 2017.

McMahon, K., et al. "Beyond Idea Generation: The Power of Groups in Developing Ideas." *Creativity Research Journal*, 28, 2016.

Microsoft. "Hendrick Motorsports Uses Microsoft Teams to Win Productivity Race." *Customers.Microsoft.com*, April 27, 2017.

Moran, C. "How Should Your Company Prepare for Robot Coworkers?" *Fast Company*, February 13, 2018.

Morar HPI. "A Global Survey Reveals Employee Perception of Advanced Technologies and Virtual Assistants in the Workplace." *Cisco.com*, October 2017.

Mulgan, G. *Big Mind: How Collective Intelligence Can Change Our World*. Princeton, NJ: Princeton University Press, 2017.

Nizri, G. "Shaping the Future of Work: A Collaboration of Humans and AI." *Forbes.com*, August 17, 2017.

Pena, S. "12 Benefits of a Collaborative Workspace." *Creator*, June 14, 2017. **wework.com/creator/start-your-business/12-benefits-of-a-collaborative-workspace/** (accessed July 2018).

Power, B. "Improve Decision-Making with Help from the Crowd." *Harvard Business Review*, April 8, 2014.

Reese, H. "New Research Shows That Swarm AI Makes More Ethical Decisions Than Individuals." *Tech Republic*, June 8, 2016.

Ruiz-Hopper, M. "Hendrick Motorsports Gains Competitive Advantage on the Race Track." *Microsoft.com*, September 26, 2016.

Staff Writers. "Scientists Simulate a Space Mission in Mars-Analogue Utah Desert." *Mars Daily*, October 19, 2016.

Stewart, C. "The 18 Best Tools for Online Collaboration." *Creative Blog*, March 7, 2017.

Thorn, C., and J. Huang. "How Carnegie Is Using Technology to Enable Collaboration in Networks." *Carnegie Foundation Blog*, September 9, 2014.

Tobe, F. "Why Co-Bots Will Be a Huge Innovation and Growth Driver for Robotics Industry." *IEEE Spectrum*, December 30, 2015.

Unanimous AI. "XPRIZE Uses Swarm AI Technology to Optimize Visioneers Summit Ideation." *Unanimous AI*, 2018. **UAI_case_study_xprize_0601_0601.pdf** (accessed July 2018).

Wladawsky-Berger, I. "Building an Effective Human-AI Decision System." *The Wall Street Journal*, December 1, 2017.

Xia, L. "Improving Group Decision-Making by Artificial Intelligence." In C. Sierra, Editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI*, 2017.

Yazdani, M., et al. "A Group Decision Making Support System in Logistics and Supply Chain Management." *Expert Systems with Applications*, 88, December 1, 2017.

Yoon, Y., et al. "Preference Clustering-Based Mediating Group Decision-Making (PCM-GDM) Method for Infrastructure Asset Management." *Expert Systems with Applications*, 83, October 15, 2017.

Yurieff, K. "Robot Predicts Boston Will Win Amazon HQ2." *CNN Tech*, March 13, 2018a.

Yurieff, K. "Robot Co-Workers? 7 Cool Technologies Changing the Way We Work." *CNN Tech*, May 4, 2018b.

CHAPTER 12

Knowledge Systems: Expert Systems, Recommenders, Chatbots, Virtual Personal Assistants, and Robo Advisors

LEARNING OBJECTIVES

- Describe recommendation systems
- Describe expert systems
- Describe chatbots
- Understand the drivers and capabilities of chatbots and their use
- Describe virtual personal assistants and their benefits
- Describe the use of chatbots as advisors
- Discuss the major issues related to the implementation of chatbots

Advancement in artificial intelligence (AI) technologies and especially natural language processing (NLP), machine and deep learning and knowledge systems, coupled with the increased quality and functionalities of other intelligent systems, and mobile devices and their apps, have driven the development of chatbots (bots) for inexpensive and fast execution of many tasks related to communication, collaboration, and information retrieval. The use of chatbots in business is increasing rapidly, partly because of their fit with mobile systems and devices. As a matter of fact, sending messages is probably the major activity in the mobile world.

In the last two to three years, many thousands of bots have been placed into service worldwide by both organizations (private and public) and individuals. Many people refer to these phenomena as the *chatbot revolution*. Chatbots today are much more sophisticated than those of the past. They are extensively used, for example, in marketing; customer, government, and financial services; healthcare; and in manufacturing. Chatbots make communication more personal than faceless computers and excel in data gathering. Chatbots can stand alone or be parts of other knowledge systems.

We divide the applications in this chapter into four categories: expert systems, chatbots for communication and collaboration, virtual personal assistants (native products, such as Alexa), and chatbots that are used as professional advisors. Some implementation topics of intelligent systems are described last.

This chapter has the following sections:

12.1 OPENING VIGNETTE: Sephora Excels with Chatbots

THE PROBLEM

Sephora is a French-based cosmetics/beauty products company doing business globally. It has its own stores and sells its goods in cosmetic and department stores. In addition, Sephora sells online on Amazon and on its online store. The company sells hundreds of brands, including many of its own. It operates in a very competitive market where customer care and advertising are critical. Sephora sells some products for men, but most beauty products are targeted to women.

THE SOLUTION

Sephora's first use of chatbots occurred through messaging services. The purpose of the first bot was to search for information for the company's resources such as videos, images, tips, and so on. This bot operates in a question-and-answer (Q&A) mode. It recommends relevant content based on customers' interests. The company aims to appeal to young customers messaging on Kik.

Sephora researchers found that customers conversing with the Kikbot were engaged deeply in the dialog. Then the bot encouraged them to explore new products. Sephora's newer bot called Reservation Assistant was placed on Facebook Messenger. It enables customers to book or reschedule makeover appointments.

Another Sephora bot delivered on Kik is Shade-Matching. It matches lips colors to photos (face and lips) uploaded by users and recommends the best match to them. The bot also lets users try on photos of recommended colors, using Sephora Virtual Artist that runs on Facebook Messenger. Bots are deployed as mobile apps. If users like the recommendation, they are directed to the company's Web store to buy the products. Users can upload photos taken with selfies so that the program can do the matching. Over 4 million visitors tried 90 million shades in the first year of Virtual Artist's operation.

The Q&A collection of the knowledge base was built by connecting it with store experts. Knowledge acquisition techniques (Chapter 2) were used for this purpose. The company's bots use NLPs that were trained to understand the typical vocabulary of users.

THE RESULTS

The company's customers loved the bots. In addition, Sephora learned the importance of providing assistance and guidance to users who are motivated to return (at a reasonable cost!), happier, and more engaged.

Sephora's bot asks users questions to find their tastes and preferences. Then it acts like a *recommendation system* (Section 12.2), offering products. Kik and Messenger users can purchase items without leaving the messaging service.

Finally, the company has improved the bots' knowledge over time and plans new bots for additional tasks.

Note: Sephora was selected by *Fast Company Magazine*, March/April 2018, as one of the "World's Most Innovative Companies." Sephora is known for its digital transformation and innovation (Rayome, 2018). Also, Sephora's bots are considered among the top marketing chatbots (Quoc, 2017).

Sources: Compiled from Arthur (2016), Rayome (2018), and Taylor (2016), **theverge.com/2017/3/16/14946086/sephora-virtual-assistant-ios-app-update-ar-makeup/**, and **sephora.com/**.

QUESTIONS FOR THE OPENING VIGNETTE

1. List and discuss the benefits of bots to the company.
2. List and discuss the benefits of bots to customers.
3. Why were the bots deployed via Messenger and Kik?
4. What would happen to Sephora if competitors use a similar approach?

WHAT WE CAN LEARN FROM THIS VIGNETTE

In the highly competitive world of retail beauty products, customer care and marketing are critical. Using only live employees can be very expensive. In addition, customers are shopping 24/7, and physical stores are open during limited hours and days. In addition, there are large combinations of certain beauty products (e.g., many shades/colors) available. Sephora decided to use chatbots on Facebook Messenger and Kik to engage its customers. Chatbots, the subject of this chapter, are available 24/7 at a lower cost and are delivered via mobile devices. Bots deliver information to customers consistently and quickly direct customers to easy online shopping. Sephora placed its chatbots on messaging services. The logic was that people like to chat with friends on messaging services, and they may also like to chat with businesses.

In addition to several services to customers, using chatbots helps Sephora learn about customers. This type of chatbot is the most common type for customer care and marketing. In this chapter, we cover several other types of knowledge systems, including the pioneering expert systems, recommenders, virtual personal assistants offered by several large technology companies, and robo advisors.

12.2 EXPERT SYSTEMS AND RECOMMENDERS

In Chapter 2 we introduced the reader to the concept of autonomous decision systems. An *expert system* is a category of autonomous decision systems and are considered the earliest applications of AI. Expert systems use started in research institutions in the early and mid-1960s (e.g., Stanford University, IBM) and was adopted commercially during the 1980s.

Basic Concepts of Expert Systems (ES)

The following are the major concepts related to ES technology.

DEFINITIONS There are several definitions of expert systems. Our working definition is that an **expert system** is a computer-based system that emulates decision making and/or problem solving of human experts. These decisions and problems are in complex areas

that require expertise to solve. The basic objective is to enable nonexperts to make decisions and solve problems that usually require expertise. This activity is usually performed in narrowly defined domains (e.g., making small loans, providing tax advice, analyzing reasons for machine failure). Classical ES use "what-if-then" rules for their reasoning.

EXPERTS An ***expert*** is a person who has the special knowledge, judgment, experience, and skills to provide sound advice and solve complex problems in a narrowly defined area. It is an expert's job to provide the knowledge about how to perform a task so that a nonexpert will be able to do the same task assisted by ES. An expert knows which facts are important and understands and explains the dependent relationships among those facts. In diagnosing a problem with an automobile's electrical system, for example, an expert car mechanic knows that a broken fan belt can be the cause for the battery to discharge.

There is no standard definition of *expert*, but decision performance and the level of knowledge a person has are typical criteria used to determine whether a particular person is an expert as related to ES. Typically, experts must be able to solve a problem and achieve a performance level that is significantly better than average. An expert at one time or in one region may not be an expert in another time or region. For example, a legal expert in New York may not be one in Beijing, China. A medical student may be an expert compared to the general public but not in making a diagnosis or performing surgery. Note that experts have expertise that can help solve problems and explain certain obscure phenomena only within a specific domain.
Typically, human experts are capable of doing the following:

- Recognizing and formulating a problem.
- Solving a problem quickly and correctly.
- Explaining a solution.
- Learning from experience.
- Restructuring knowledge.
- Breaking rules (i.e., going outside the general norms) if necessary.
- Determining relevance and associations.

Can a machine help a nonexpert perform like an expert? Can a machine make autonomous decisions that experts make? Let us see. But first, we need to explore what expertise is.

EXPERTISE An ***expertise*** is the extensive, task-specific knowledge that experts possess. The level of expertise determines the success of a decision made by an expert. Expertise is often acquired through training, learning, and experience in practice. It includes explicit knowledge, such as theories learned from a textbook or a classroom and implicit knowledge gained from experience. The following is a list of possible knowledge types used in ES applications:

- Theories about the problem domain.
- Rules and procedures regarding the general problem domain.
- Heuristics about what to do in a given problem situation.
- Global strategies for solving of problems amenable to expert systems.
- Meta knowledge (i.e., knowledge about knowledge).
- Facts about the problem area.

These types of knowledge enable experts to make better and faster decisions than nonexperts.

Expertise often includes the following characteristics:

- It is usually associated with a high degree of intelligence, but it is not always associated with the smartest person.
- It is usually associated with a vast quantity of knowledge.
- It is based on learning from past successes and mistakes.
- It is based on knowledge that is well stored, organized, and quickly retrievable from an expert who has excellent recall of patterns from previous experiences.

Characteristics and Benefits of ES

ES were used during the period 1980 to 2010 by hundreds of companies worldwide. However, since 2011, their use has declined rapidly, mostly due to the emergence of better knowledge systems, three types of which are described in this chapter. It is important, however, to understand the major characteristics and benefits of expert systems since many of them evolved evidenced newer knowledge systems.

The major objective of ES is the transfer of expertise to a machine. The expertise will be used by nonexperts. A typical example is a diagnosis. For example, many of us can use self-diagnosis to find (and correct) problems in our computers. Even more than that, computers can find and correct problems by themselves. One field in which such ability is practiced is medicine, as described in the following example:

Example: Are You Crazy?

A Web-based ES was developed in Korea for people to self-check their mental health status. Anyone in the world can access it and get a free evaluation. The knowledge for the system was collected from a survey of 3,235 Korean immigrants. The results of the survey were analyzed and then reviewed by experts via focus group discussions. For more information, see Bae (2013).

BENEFITS OF ES Depending on the mission and structure of ES, the following are their capabilities and potential benefits:

- Perform routine tasks (e.g., diagnosis, candidate screening, credit analysis) that require expertise much faster than humans.
- Reduce the cost of operations.
- Improve consistency and quality of work (e.g., reduce human errors).
- Speed up decision making and make consistent decisions.
- May motivate employees to increase productivity.
- Preserve scarce expertise of retiring employees.
- Help transfer and reuse knowledge.
- Reduce employee training cost by using self-training.
- Solve complex problems without experts and solve them faster.
- See things that even experts sometimes miss.
- Combine expertise of several experts.
- Centralize decision making (e.g., by using the "cloud").
- Facilitate knowledge sharing.

These benefits can provide a significant competitive advantage to companies that use ES. Indeed, some companies have saved considerable amounts of money using them.

Despite these benefits, the use of ES is on the decline. The reasons for this and the related limitations are discussed later in this section.

Typical Areas for ES Applications

ES have been applied commercially in a number of areas, including the following:

- ***Finance.*** Finance ES include analysis of investments, credit, and financial reports; evaluation of insurance and performance; tax planning; fraud prevention; and financial planning.
- ***Data processing.*** Data processing ES include system planning, equipment selection, equipment maintenance, vendor evaluation, and network management.
- ***Marketing.*** Marketing ES include customer relationship management, market research and analysis, product planning, and market planning. Also, presale advice is provided for prospects.
- ***Human resources.*** Examples of human resource ES are planning, performance evaluation, staff scheduling, pension management, regulatory advising, and design of questionnaires.
- ***Manufacturing.*** Manufacturing ES include production planning, complex product configuration, quality management, product design, plant site selection, and equipment maintenance and repair (including diagnosis).
- ***Homeland security.*** These ES include terrorist threat assessment and terrorist finance detection.
- ***Business process automation.*** ES have been developed for desk automation, call center management, and regulation enforcement.
- ***Healthcare management.*** ES have been developed for bioinformatics and other healthcare management issues.
- ***Regulatory and compliance requirements.*** Regulations can be complex. ES are using a stepwise process to ensure compliance.
- ***Web site design.*** A good Web site design requires paying attention to many variables and ensures that performance is up to standard. ES can lead to a proper design process.

Now that you are familiar with the basic concepts of ES, it is time to look at the internal structure of ES and how their goals are achieved.

Structure and Process of ES

As you may recall from Section 2.5 and Figure 2.5, the process of knowledge extraction and its use is divided into two distinct parts. In ES we refer to these as the *development environment* and the *consultation environment* (see Figure 12.1). An ES builder builds the necessary ES components and loads the knowledge base with appropriate representation of expert knowledge in the **development environment**. A nonexpert uses the **consultation environment** to obtain advice and solve problems using the expert knowledge embedded into the system. These two environments are usually separated.

MAJOR COMPONENTS OF ES The major components in typical expert systems include:

- *Knowledge acquisition.* Mostly from human experts, is usually obtained by knowledge engineers. This knowledge, which may derive from several sources, is integrated, validated, and verified.
- *Knowledge base.* This is a knowledge repository. The knowledge is divided into knowledge about the domain and knowledge about problem solving and solution procedures. Also, the input data provided by the users may be stored in the knowledge base.
- *Knowledge representation.* This is frequently organized as business rules (also known as *production rules*).

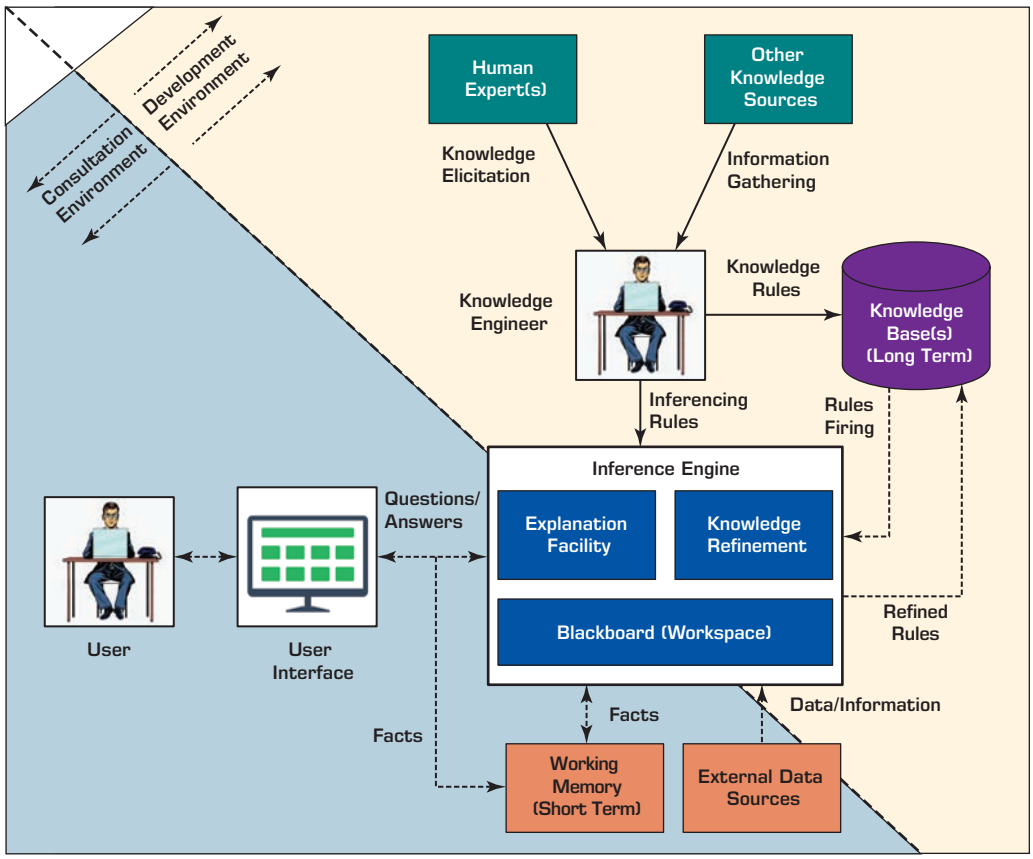


FIGURE 12.1 General Architecture of Expert Systems.

- *Inference engine.* Also known as the *control structure* or the *rule interpreter*, this is the "brain" of ES. It provides the reasoning capability, namely the ability to answer users' questions, provide recommendations for solutions, generate predictions, and conduct other relevant tasks. The engine manipulates the rules by either forward chaining or backward chaining. In 1990s ES started to use other inference methods.
- *User interface.* This component allows user inference engine interactions. In classical ES, this was done in writing or by using menus. In today's knowledge systems, it is done by natural languages and voice.

These major components of ES generate useful solutions in many areas. Remember that these areas need to be well structured and in fairly narrow domains. Less common is a *justifier/explanation* subsystem that shows users of rule-based systems the chains of rules used to arrive at conclusions. Also, least common is a *knowledge refining subsystem* that helped to improve knowledge (e.g., rules) when new knowledge is added.

A major provider of expert systems technologies was Exsys Inc. While the company is no longer active in this business, its Web site (**Exsys.com**) is. It contains tutorials and a large number of cases related to its major software product, *Exsys Corvid*. Application Case 12.1 is one example.

Application Case 12.1

ES Aid in Identification of Chemical, Biological, and Radiological Agents

Terrorist attacks using chemical, biological, or radiological (CBR) agents are of great concern due to their potential for leading to large loss of life. The United States and other nations have spent billions of dollars on plans and protocols to defend against acts of terrorism that could involve CBR. However, CBR covers a wide range of input agents with many specific organisms that could be used in multiple ways. Timely response to such attacks requires rapid identification of the input agents involved. This can be a difficult process involving different methods and instruments.

The U.S. Environmental Protection Agency (EPA) along with Dr. Lawrence H. Keith, president of Instant Reference Sources Inc. and other consultants, have incorporated their knowledge, experience, and expertise as well as information in publicly available EPA documents to develop the CBR Advisor using Exsys Inc.'s Corvid software.

One of the most important parts of the CBR Advisor is providing advice in logical step-by-step procedures to determine the identity of a toxic agent when little or no information is available, which is typical at the beginning of a terrorist attack. The system helps response staff proceed according to a well-established action plan even in such a highly stressful environment. The system's dual screens present three levels of information: (1) a top/executive level with brief answers, (2) an educational level with in-depth information, and (3) a research level with links to other documents, slide shows, forms, and Internet sites. CBR Advisor's content includes:

- How to classify threat warnings.
- How to conduct an initial threat evaluation.
- What immediate response actions to take.
- How to perform site characterization.
- How to evaluate the initial site and safe entry to it.
- Where and how to best collect samples.
- How to package and ship samples for analysis.

Restricted content includes CBR agents and methods for analyzing them. The CBR Advisor can be used for incident response and/or training. It has two different menus, one for emergency response and another, longer menu for training. It is a restricted software program and is not publicly available.

Questions for Case 12.1

1. How can the CBR Advisor assist in making quick decisions?
2. What characteristics of the CBR Advisor make it an expert system?
3. What could be other situations in which similar expert systems can be employed?

Expert systems are also used in high-pressure situations in which human decision makers often need to take split-second actions involving both subjective as well as objective knowledge in responding to emergency situations.

Sources: **www.exsys.com** "Identification of Chemical, Biological and Radiological Agents" **http://www.exsyssoftware.com/CaseStudySelector/casestudies.html**. April 2018. (Publicly available information.) Used with permission.

Why the Classical Type of ES Is Disappearing

The large benefits described earlier drove the implementation of many ES worldwide. However, like many other technologies, the classical ES have been replaced by better systems. Let us first look at some of the limitations of ES that contributed to its declining use.

1. The acquisition of knowledge from human experts has proven to be very expensive due to the shortage of good knowledge engineers as well as the possible need to interview several experts for one application.
2. Any acquired knowledge needed to be updated frequently at a high cost.
3. The rule-based foundation was frequently not robust and not too reliable or flexible and could have too many exceptions to the rules. Improved knowledge systems use

data-driven and statistical approaches to make the inferences with better success. In addition, case-based reasoning could work better only if a sufficient number of similar cases were available. So, usually it cannot support ES.

4. The rule-based user-interface needed to be supplemented (e.g., by voice communication, image maps). This could make ES too cumbersome.
5. The reasoning capability of rule-based technology is limited compared to use of newer mechanisms such as those used in machine learning.

NEW GENERATION OF EXPERT SYSTEMS Instead of using the old knowledge acquisition and representation system, newer ES based on machine learning algorithms and other AI technologies are deployed to create better systems. An example is provided in Application Case 12.2.

Application Case 12.2 VisiRule

VisiRule is an older ES company that remodeled its business over time. VisiRule (of the United Kingdom) provides easy-to-use diagramming tools to facilitate the construction of ES. Diagramming allows easier extraction and use of knowledge in expert systems.

The process of building the knowledge base can be seen on the left side of Figure 12.2. On the left-hand side, you can see the hybrid creation. Using a decision tree, the domain experts can create additional rules directly from relevant data (e.g., historical). In addition, rules can be created by machine learning (lower left side).

The right-hand side (upper corner) illustrates the hybrid delivery (consultation). Using interactive questions and answers the system can generate advice. In addition, rules can be used to process data remotely and update the data repository. Note that the dual delivery option is based on machine learning's ability to discover hidden patterns in data that can be used to form predictive decision models.

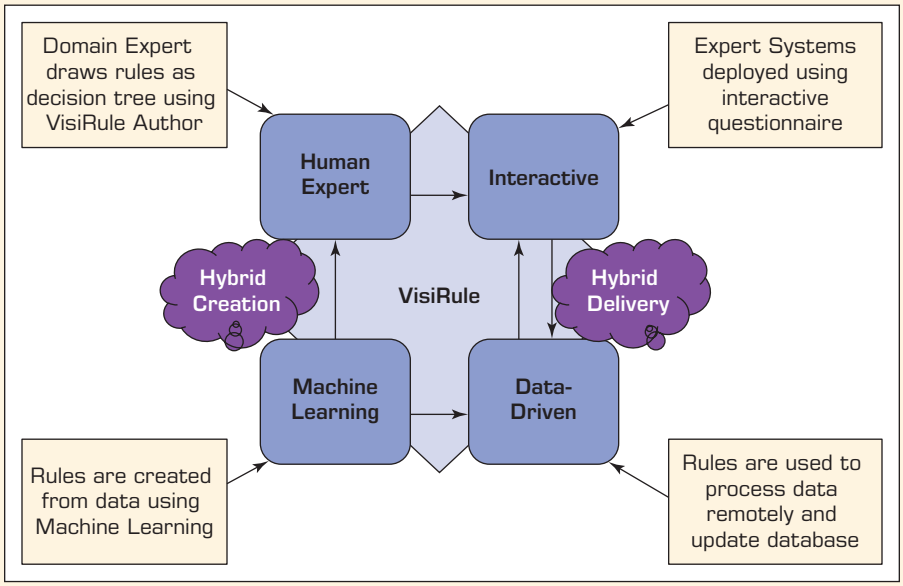


FIGURE 12.2 The Process of Recommendation Systems.

VisiRule also provides chatbots for improving the interactive part of the process and supplies an interactive map. According to the company's Web site **visirule.co.uk/**, the major benefits of the product are:

- It is code-free; no programming is needed.
- The diagrams are drawn by human experts or induced automatically from data.
- It contains self-assessment tools with report generation and document production.
- The generated knowledge can be easily executed as XML code.
- It provides explanation and justification.
- The interactive expert advice attracts new customers.
- It can be used for training and advising employees.
- Companies can easily access the corporate knowledge repository.
- The charts to use VisiRule authoring tools are created with ease using flowcharting and decision trees.
- The charts allow creation of models that can be immediately executed and validated.

All-in-all, VisiRule provides a comprehensive AI-based expert system.

Source: Courtesy of VisiRule Corp. UK. Used with permission.

Questions for Case 12.2

1. Which of the limitations of early ES have been solved by the VisiRule system?
2. Compare Figures 12.2 and 12.1. What are the differences between the creation (Fig. 12.2) and the development (Fig. 12.1) subsystems?
3. Compare Figures 12.2 and 12.1. What are the differences between the delivery (Fig. 12.2) and the consultation (Fig. 12.1) subsystems?
4. Identify all AI technologies and list their contribution to the VisiRule system.
5. List some benefits of this ES to users.

Three major AI types of applications that overcome the earlier discussed limitations of RS are chatbots, virtual personal assistants, and robo advisors, which are presented next in this chapter. Other AI technologies that perform similar activities are presented in Chapters 4 to 9. Most notable is IBM Watson (Chapter 6); some of its advising capabilities are similar to those of ES but are much superior.

Another similar AI technology, the recommendation system, is presented next. Its newer variations use machine learning and IBM Watson Analytics.

Recommendation Systems

A heavily used knowledge system for recommending one-to-one targeted products or services is the **recommendation system**, also known as *recommender system* or *recommendation engine*. Such a system tries to predict the importance (rating or preference) that a user will attach to a product or service. Once the rating is known, a vendor knows users' tastes and preferences and can match and recommend a product or service to the user. For comprehensive coverage, see Aggarwal (2016). For a comprehensive tutorial and case study, see **analyticsvidhya.com/blood/2015/10/recommendation-engines/**.

Recommendation systems are very common and are used in many areas. Top applications include *movies, music*, and *books*. However, there are also systems for *travel, restaurants, insurance*, and *online dating*. The recommendations are typically given in rank order. Online recommendations are preferred by many people over regular searches, which are less personalized, slower, and sometimes less accurate.

BENEFITS OF RECOMMENDATION SYSTEMS Using these systems may result in substantial benefits both to buyers and sellers (see Makadia, 2018).

Benefits to customers are:

- *Personalization.* They receive recommendations that are very close to fulfilling what they like or need. This depends, of course, on the quality of the method used.
- *Discovery.* They may receive recommendations for products that they did not even know existed but were what they really need.
- *Customer satisfaction.* With repeated recommendations tends to increase.
- *Reports.* Some recommenders provide reports and others provide explanations about the selected products.
- *Increased dialog with sellers.* Because recommendations may come with explanations, buyers may want more interactions with the sellers.

Benefits to sellers are:

- *Higher conversion rate.* With personalized product recommendations, buyers tend to buy more.
- *Increased cross-sell.* Recommendation systems can suggest additional products. **Amazon.com**, for example, shows other products that "people bought together with the product you ordered."
- *Increased customer loyalty.* As benefits to customers increase, their loyalty to the seller increases.
- *Enabling of mass customization.* This provides more information on potential customized orders.

Several methods are (or were) used for building recommendation systems. Two classic methods are *collaborative filtering* and *content-based filtering*.

COLLABORATIVE FILTERING This method builds a model that summarizes the past behavior of shoppers, how they surf the Internet, what they were looking for, what they have purchased, and how much they like (rate) the products. Furthermore, collaborative filtering considers what shoppers with similar profiles bought and how they rated their purchases. From this, the method uses AI algorithms to *predict* the preference of both old and new customers. Then, the computer program makes a recommendation.

CONTENT-BASED FILTERING This technique allows vendors to identify preferences by the attributes of the product(s) that customers have bought or intend to buy. Knowing these preferences, the vendor recommends to customers products with similar attributes. For instance, the system may recommend a text-mining book to a customer who has shown interest in data mining, or action movies after a consumer has rented one in this category.

Each of these types has advantages and limitations (see example at **en.wikipedia.org/wiki/Recommender_system**). Sometimes the two are combined into a unified method.

Several other filtering methods exist. Examples include rule-based filtering and activity-based filtering. Newer methods include machine learning and other AI technologies, as illustrated in Application Case 12.3.

Application Case 12.3 Netflix Recommender: A Critical Success Factor

According to **ir.netflix.com**, Netflix is (Spring 2018 data) the world's leading Internet television network with more than 118 million members in over 190 countries enjoying more than 150 million hours of TV shows and movies per day, including original series, documentaries, and feature films. Members can view unlimited shows without commercials for a monthly fee.

The Challenges

Netflix has several million titles and now produces its own shows. The large titles inventory often creates a problem for customers who have difficulty determining which offerings they want to watch. An additional challenge is that Netflix expanded its business from the United States and Canada to 190 other countries. Netflix operates in a very competitive environment in which large players such as Apple, **Amazon.com**, and Google operate. Netflix was looking for a way to distinguish itself from the competition by making useful recommendations to its customers.

The Original Recommendation Engine

Netflix originally was solely a mail-order business for DVDs. At that time, it encountered inventory problems due to its customers' difficulties in determining which DVDs to rent. The solution was to develop a recommendation engine (called Cinematch) that told subscribers which titles they probably would like. Cinematch used data mining tools to sift through a database of billions of film ratings and customers' rental histories. Using proprietary algorithms, it recommended rentals to customers. The recommendation was accomplished by comparing an individual's likes, dislikes, and preferences against those of people with similar tastes, using a variant of *collaborative filtering*. Cinematch was like the geeky clerk at a small movie store who sets aside titles he knows you will like and suggests them to you when you visit the store.

To improve Cinematch's accuracy, Netflix began a contest in October 2016, offering $1 million to the first person or team that will write a program that would increase Cinematch's prediction accuracy by at least 10 percent. The company understood that this would take quite some time; therefore, it offered a $50,000 Progress Prize each year in which the contest was conducted. After more than two years of competition, the grand prize went to Bellkor's Pragmatic Chaos team, a combination of two runner-up teams.

To learn how the movie recommendation algorithms work, see **quora.com/How-does-the-Netflix-movie-recommendation-algorithm-work/**.

The New Era

As time passed, Netflix moved to the streaming business and then to Internet TV. Also, the spread of *cloud technology* enabled improvement in the recommendation system. The new system stopped making recommendations based on what *people have seen in the past*. Instead, it is using Amazon's cloud to mimic the human brain in order to find what people really like in their favorite movies and shows. The system is based on AI and its technology of *deep learning*. The company can now visualize Big Data and draw insights for the recommendations. The analysis is also used in creating the company's productions. Another major change dealt with the transformation to the global arena. In the past, recommendations had been based on information collected in the country (or region) where users live. The recommendations were based on what other people in the same country enjoyed. This approach did not work well in the global environment due to cultural, political, and social differences. The modified system considers what people who live in many countries view and their viewing habits and likes.

Implementation of the new system was difficult, especially when a new country or region was added. Recommendations were initially made without knowing much about the new customers. It took 70 engineers and a year of work to modify the recommendation system. For details, see Popper (2016).

The Results

As a result of implementing its recommender system, Netflix has seen very fast growth in sales and membership. The benefits include the following:

- **Effective recommendations.** Many Netflix members select their movies based on recommendations tailored to their individual tastes.
- **Customer satisfaction.** More than 90 percent of Netflix members say they are so satisfied with the Netflix service that they recommend it to family members and friends.
- **Finance.** The number of Netflix members has grown from 10 million in 2008 to 118 million in 2018. Its sales and profits are climbing steadily. In spring 2018, Netflix stock sold for over $400 per share compared with $140 a year earlier.

Sources: Based on Popper (2016), Arora (2016), and StartUp (2016).

(*Continued*)

Application Case 12.3 (Continued)

Questions for Case 12.3

1. Why is the recommender system useful? (Relate it to one-to-one targeted marketing.)
2. Explain how recommendations are generated.
3. Amazon disclosed its recommendation algorithms to the public but Netflix did not. Why?
4. Research the research activities that attempt to "mimic the human brain."
5. Explain the changes due to the globalization of the company.

SECTION 12.2 REVIEW QUESTIONS

1. Define *expert systems*.
2. What is the major objective of ES?
3. Describe experts.
4. What is expertise?
5. List some areas especially amenable to ES.
6. List the major components of ES and describe each briefly.
7. Why is ES usage on the decline?
8. Define *recommendation systems* and describe their operations and benefits.
9. How do recommendation systems relate to AI?

12.3 CONCEPTS, DRIVERS, AND BENEFITS OF CHATBOTS

The world is now infested with chatbots. According to 2017 data (Knight, 2017c), 60 percent of millennials have already used chatbots and 53 percent of those who have not used them are interested in doing so. Millennials are not the only generation using chatbots, although they may use them more than others. What chatbots are and what they do is the subject of this section.

What Is a Chatbot?

Short for *chat robot*, a **chatbot**, also known as a "bot" or "robo," is a computerized service that enables easy conversations between humans and humanlike computerized robots or image characters, sometimes over the Internet. The conversations can be in writing, and more and more are by voice and images. The conversations frequently involve short questions and answers and are executed in a natural language. More intelligent chatbots are equipped with NLPs, so the computer can understand unstructured dialog. Interactions also can occur by taking or uploading images (e.g., as is done by Samsung Bixby on the Samsung S8 and 8). Some companies experiment with *learning chatbots*, which gain more knowledge with their accumulated experience. The ability of the computer to converse with a human is provided by a knowledge system (e.g., rule-based) and a natural language understanding capability. The service is often available on *messaging services* such as Facebook Messenger or WeChat, and on Twitter.

Chatbot Evolution

Chatbots originated decades ago. They were simple ES that enabled machines to answer questions posted by users. The first known such machine was Eliza (**en.wikipedia.org/wiki/ELIZA**). Eliza and similar machines were developed to work in Q&A mode.

The machine evaluated each question, usually to be found in a bank of FAQs, and generated an answer matched to each question. Obviously, if the question was not in the FAQ collection, the machine provided irrelevant answers. In addition, because the power of the natural language understanding was limited, some questions were misunderstood and the answers were at times at best entertaining. Therefore, many companies opted to use live chats, some with inexpensive labor, organized as call centers around the globe. For more about Eliza's current generation, and how to build it, see **search.cpan.org/dist/Chatbot-Eliza/Chatbot/Eliza.pm/**. Chatbot use and reputation are rapidly increasing globally.

Example

Sophia is a chatbot created in Hong Kong and was awarded citizenship by Saudi Arabia in October 2017. Because she is not a Muslim, she is not wearing a hijab. She can answer many questions. For details, see **newsweek.com/Saudi-arabia-robot-sophia-muslim-694152/**.

TYPES OF BOTS Bots can be classified by their capabilities; three classes follow:

1. ***Regular bots.*** These are essentially conversational intelligent agents (Chapter 2). They can do simple, usually repetitive, tasks for their owners, such as showing their bank's debits, helping them to purchase goods online, and to sell or buy stocks online.
2. ***Chatbots.*** In this category, we include more capable bots, for example, those that can stimulate conversations with people. This chapter deals mainly with chatbots.
3. ***Intelligent bots.*** These have a knowledge base that is improving with experience. That is, these bots can learn, for example, a customer's preferences (e.g., like Alexa and some robo advisors).

A major limitation of the older types of bots was that updating their knowledge base was both slow and expensive. They were developed for specific narrow domains and/or specific users. It took many years to improve the supporting technology. NLP has become better and better. Knowledge bases are updated today in the "cloud" in a central location; the knowledge is shared by many users so the cost per user is reduced.

The stored knowledge is matched with questions asked by users. The answers by the machines have improved dramatically. Since 2000, we have seen more and more capable AI machines for Q&A dialogs. Around 2010, conversational AI machines were named chatbots and later were developed into virtual personal assistants, championed by Amazon's Alexa.

DRIVERS OF CHATBOTS The major drivers are:

- Developers are creating powerful tools to build chatbots quickly and inexpensively with useful functionalities.
- The quality of chatbots is improving, so conversations are getting more useful to users.
- Demand for chatbots is growing due to their potential cost reduction and improved customer service and marketing services, which are provided 24/7.
- Use of chatbots allows rapid growth without the need to hire and train many customer service employees.
- Using chatbots, companies can utilize the messaging systems and related apps that are the darlings for consumers, especially younger ones.

Components of Chatbots and the Process of Their Use

The major components of chatbots are:

- A person (client).
- A computer, avatar, or robot (the AI machine).
- A knowledge base that can be embedded in the machine or available and connected to the "cloud."
- A human-computer interface that provides the dialog for written or voice modes.
- An NLP that enables the machine to understand natural language.

Advanced chatbots can also understand human gestures, cues, and voice variations.

PERSON-MACHINE INTERACTION PROCESS The components just listed provide the framework for people-bot conversation. Figure 12.3 shows the conversation process.

- A person (left side of the figure) needs to find some information, or need some help.
- The person asks a related question from the bot by voice, texting, and so on.
- NLP translates the question to machine language.
- The chatbot transfers the question to cloud services.
- The cloud contains a knowledge base, business logic, and analytics (if appropriate) to craft a response to the question.
- The response is transferred to a natural language generation program and then to the person who asked the question in the preferred mode of dialog.

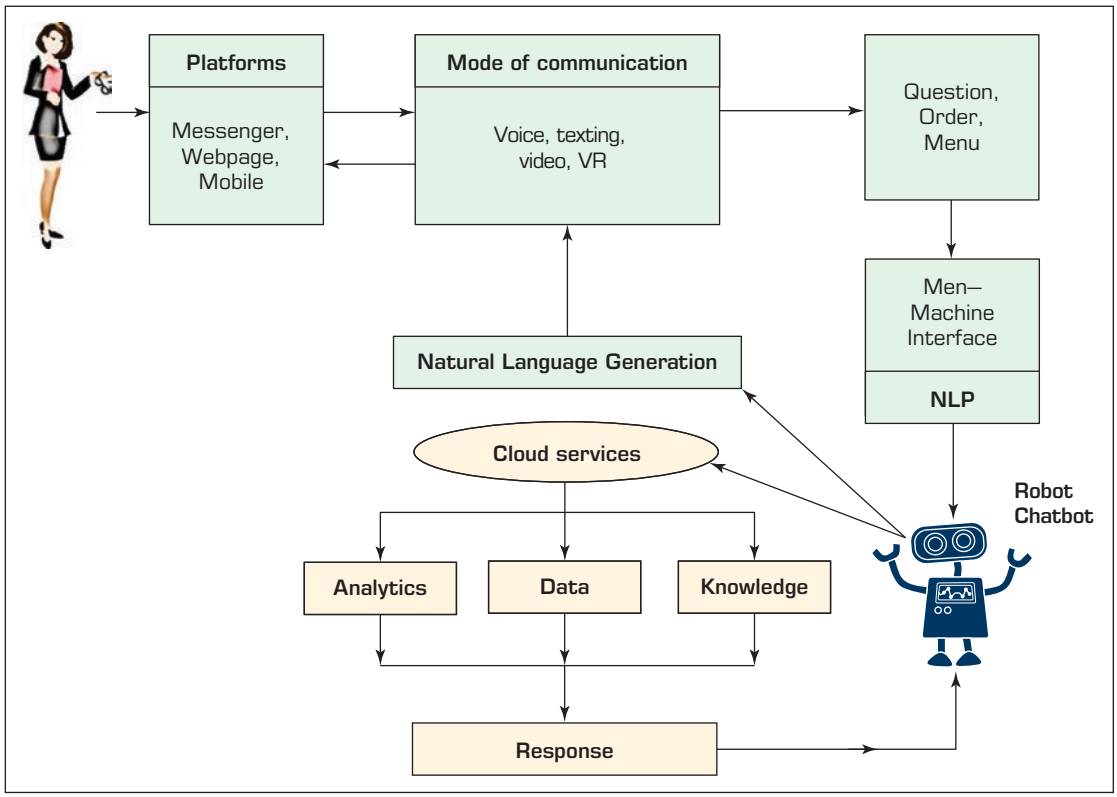


FIGURE 12.3 The Process of Chatting with Chatbots.

Drivers and Benefits

Chatbot use is driven by the following forces and benefits:

- The need to cut costs.
- The increasing capabilities of AI, especially NLP and voice technologies.
- The ability to converse in different languages (via machine translation).
- The increased quality and capability of captured knowledge.
- The push of devices by vendors (e.g., virtual personal assistants such as Alexa from Amazon and Google Assistant from Alphabet).
- Its use for providing superb and economic customer service and conducting market research.
- Its use for text and image recognition.
- Its use to facilitate shopping.
- Its support of decision making.

Chatbots and similar AI machines have been improved over time. Chatbots are beneficial to both users and organizations. For example, several hospitals employ robot receptionists to direct patients to their place of treatment. Zora Robotics created a robot named Nao to act as a chatting companion for people who are sick or elderly. The bot acts, for example, as a form of therapy for those suffering from dementia.

Note: For some limitations of chatbots, see Section 12.7.

Representative Chatbots from Around the World

For a chatbot directory of the more than 1,250 bots in 53 countries as of April 2018, see **chatbots.org/** and at **botlist.co/bots/**. Examples of chatbots and what they can do from **chatbot.org/** are provided here:

- **RoboCoke.** This is a party and music recommendation bot created for Coca-Cola in Hungary.
- **Kip.** This shopping helper is available on Slack (a messaging platform). Tell Kip what you want to buy, and Kip will find it and even buy it for you.
- **Walnut.** This chatbot can discover skills relevant to you and help you learn them. It analyzes a large set of data points to discover the skills.
- **Ride sharing by Taxi Bot.** If you are not sure whether Uber, Lyft, Grab, or Comfort DelGro is the cheapest service, you can ask this bot. In addition, you can get current promo codes.
- **ShopiiBot.** When you send a picture of a product to this bot, it will find similar ones in seconds. Alternatively, tell ShopiiBot what kind of product you are looking for at what price, and it will find the best one for you.
- **Concerning desired trips.** It can answer questions regarding events, restaurants, and attractions in major destinations.
- **BO.T.** The first Bolivian chatbot, it talks to you (in Spanish) and answers your questions about Bolivia, its culture, geography, society, and more.
- **Hazie.** She is your digital assistant that aims to close the gap between you and your next career move. Job seekers can converse directly with Hazie just as they do with a job placement agent or friends.
- **Green Card.** This Visabot product helps users to properly file requests for Green Cards in the United States.
- **Zoom.** Zoom.ai (**botlist.co/bots/369-zoomai**), an automated virtual assistant, is for everyone in the workplace.
- **Akita.** This chatbot (**botlist.co/bots/1314-akita**) can connect you to businesses in your area.

As you can see, chatbots can be used for many different tasks. Morgan (2017) classifies bots into the following categories: education, banking, insurance, retail, travel, healthcare, and customer experience.

MAJOR CATEGORIES OF CHATBOTS' APPLICATIONS Chatbots are used today for many purposes and in many industries and countries. We divide the applications into the following categories:

- Chatbots for enterprise activities, including communication, collaboration, customer service, and sales (such as in the opening vignette). These are described in Section 12.4.
- Chatbots that act as personal assistants. These are presented in Section 12.5.
- Chatbots that act as advisors, mostly on finance-related topics (Section 12.6).

For a discussion of these categories, see Ferron (2017).

SECTION 12.3 REVIEW QUESTIONS

1. Define *chatbots* and describe their use.
2. List the major components of chatbots.
3. What are the major drivers of chatbot technology?
4. How do chatbots work?
5. Why are chatbots considered AI machines?

12.4 ENTERPRISE CHATBOTS

Chatbots play a major role in enterprises, both in external and internal applications. Some believe that chatbots can fundamentally change the way that business is done.

The Interest of Enterprises in Chatbots

The benefits of chatbots to enterprises are increasing rapidly, making dialog less expensive and more consistent. Chatbots can interact with customers and business partners more efficiently, are available anytime, and can be reached from anywhere. Businesses are clearly paying attention to the chatbot revolution. According to Beaver (2016), businesses should look at enterprise bots for the following reasons:

- "AI has reached a stage in which chatbots can have increasingly engaging and human conversations, allowing businesses to leverage the inexpensive and wide-reaching technology to engage with more consumers.
- Chatbots are particularly well suited for mobile-perhaps more so than apps. Messaging is at the heart of the mobile experience, as the rapid adoption of chat app demonstrates.
- The chatbot ecosystem is already robust, encompassing many different third-party chat bots, native bots, distribution channels, and enabling technology companies.
- Chatbots could be lucrative for messaging apps and the developers who build bots for these platforms, similar to how app stores have developed into moneymaking ecosystems."

A study conducted in 2016 found that 80 percent of businesses want chatbots by 2020 **businessinsider.com/80-of-businesses-want-chatbots-by-2020-2016-12**. For more opportunities in marketing, see Knight (2017a).

Enterprise Chatbots: Marketing and Customer Experience

As we saw in the opening vignette to this chapter and will see in in the several examples later in this chapter, chatbots are very useful in providing marketing and customer service (e.g., Mah, 2016), obtaining sales leads, persuading customers to buy products and services, providing critical information to potential buyers, optimizing advertising campaigns (e.g., a bot named Baroj; see Radu, 2016), and much more. Customers want to do business on the app they are already in. For this reason, many bots are on Facebook Messenger, Snapchat, WhatsApp, Kik, and WeChat. Using voice and texting, it is possible to provide personalization as well as superb customer experience. Chatbots can enable vendors to improve personal relationships with customers.

In addition to the marketing areas, plenty of chatbots are in areas such as financial (e.g., banks) and HRM services as well as production and operation management for communication, collaboration, and other external and internal enterprise business processes. In general, enterprises use chatbots on messaging platforms to run marketing campaigns (e.g., see the opening vignette) and to provide superb customer experience.

IMPROVING THE CUSTOMER EXPERIENCE Enterprise chatbots create improved customer experience by providing a conversation platform for quick and 24/7 contact with enterprises. When customers benefit from the system, they are more inclined to buy and promote a specific brand. Chatbots can also supplement humans in providing improved customer experience.

EXAMPLES OF ENTERPRISE CHATBOTS Schlicht (2016) provides a beginner's guide to chatbots. He presents the following hypothetical example about today's shopping at Nordstrom (a large department store) versus the use of chatbots.

> If you wanted to buy shoes from Nordstrom online, you would go to their Web site, look around until you find the shoes you wanted, and then you would purchase them. If Nordstrom makes a bot, which I am sure they will, you would simply be able to message Nordstrom on Facebook. It would ask you what you are looking for and you would simply . . . tell it.
>
> Instead of browsing a Web site, you will have a conversation with the Nordstrom bot, mirroring the type of experience you would get when you go into the retail store.

Three additional examples follow:

Example 1: LinkedIn

LinkedIn is introducing chatbots that conduct tasks such as comparing the calendars of people participating in meetings and suggesting meeting times and places. For details, see CBS News (2016).

Example 2: Mastercard

Mastercard has two bots based on massaging platforms, one bot for banks and another bot for merchants.

Example 3: Coca-Cola

Customers worldwide can chat with Coca-Cola bots via Facebook Messenger. The bots make users feel good with conversations that are increasingly becoming personalized. The bots collect customers' data, including their interests, problems, local dialect, and attitudes and then can target advertisements tailored to each user.

A 5-min. video about Facebook is available at **cnbc.com/2016/04/13/why-facebook-is-going-all-in-on-chatbots.html**. It provides a Q&A session with David Marcus describing Facebook's increasing interest in chatbots.

WHY USE MESSAGING SERVICES? So far, we have noted that enterprises are using messaging services such as Facebook Messenger, WeChat, Kik, Skype, and WhatsApp. The reason is that in 2017, more than 2.6 billion people were chatting on messaging services. Messaging is becoming the most widespread digital behavior. WeChat of China was the first to commercialize its service by offering "chat with business" capabilities as illustrated in Application Case 12.4.

FACEBOOK'S CHATBOTS Following the example of WeChat, Facebook launched users' conversations with businesses's chatbots on a large scale on Messenger, suggesting that users could message a business just the way they would message a friend. The service allows businesses to conduct text exchanges with users. In addition, the bots have a

Application Case 12.4 WeChat's Super Chatbot

WeChat is a very large comprehensive messaging service in China and other countries with about 1 billion members in early 2018. It pioneered the use of bots in 2013 (see **mp.weixin.qq.com**). Users can use the chatbot for activities such as the following:

- Hail a taxi.
- Order food to be delivered.
- Buy movie tickets and other items.
- Customize and order a pair of Nikes.
- Send an order to the nearest Starbucks.
- Track your daily fitness progress.
- Shop Burberry's latest collection.
- Book doctor appointments.
- Pay your water bill.
- Host a business conference call.
- Send voice messages, emoticons, and snapshots to friends.
- Send voice messages to communicate with businesses.
- Communicate and engage with customers.
- Provide a framework for teamwork and collaboration.
- Conduct market research.
- Get information and recommendations on products and services.
- Launch a start-up on WeChat (you can make your own bot on WeChat for this purpose).

Griffiths (2016) has provided information concerning a Chinese online fashion flash sales company, Meici. The company used its WeChat account to gather information related to sales. Each time new users followed Meici's account, a welcome message instructed them on how to trigger resources. WeChat is available in English and other languages worldwide due to its usefulness. Facebook installed similar capabilities in 2015.

Questions for Case 12.4

1. Find some recent activities that WeChat does.
2. What makes this chatbot so unique?
3. Compare the bot of WeChat to bots offered by Facebook.

learning ability that enables them to accurately analyze people's input and provide correct responses. Overall, as of early 2018, there were more than 30,000 company bots on Facebook Messenger. Some companies use Messenger bots to recognize faces in pictures, suggesting recipients for targeted ads. According to Guynn (2016), Facebook allows software developers access to its tools that build its personal assistant called "M," which combines AI with a human touch for tasks such as ordering food or sending flowers. Using the M tools, developers can build applications for Messenger that can have an increased understanding of requests made in natural languages. A major benefit of these bots for Facebook is their collection of data and creation of profiles of users.

The following is another example of how the use of chatbots is facilitating customer service and marketing (Application Case 12.5).

Application Case 12.5 How Vera Gold Mark Uses Chatbots to Increase Sales

Vera Gold Mark is a real estate developer of luxury high rises in Punjab, India.

The Problem

Vera Gold Mark **(**VGM) is active in a very competitive market. As a developer of luxury apartments, which are usually expensive, it must try to attract many potential buyers and thus needs as many sales leads as possible at a reasonable cost. Chatting live with potential customers can be expensive since it requires very knowledgeable and courteous agents available 24/7. VGM has a large inventory of units that must be sold as soon as possible.

The Solution

VGM decided to use chatbots to supplement or replace expensive manual live chats. These work in the following ways. Buyers may click on the "chat with the robot" button on the company's Facebook page, and receive any information they need. The chat helps VGM promote its available products. When they click, users are able to chat and get information about pricing, delivery dates, construction sites, and much more for VGM projects. Users can also tweet. The chatbots provide answers about the projects. Facebook provides VGM access to potential buyers' profiles (with users' permission), which VGM sales teams can use to refine sales strategies. The system is available 24/7. Voice communication is coming soon (2018).

The Results

VGM is now viewed in a very positive way and is considered to be very professional. VGM is getting good reviews for its customer service. The builder is considered more honest and unbiased because it provides written answers and promises to customers. Salespeople at VGM get an increased number of sales leads, and because they know more about prospective customers, they can better align them with units (optimal fit). The system is also able to attract international buyers without increasing cost. Because the system is available 24/7, global buyers can easily evaluate VGM's available condominiums.

The chatbot is also used as a teaching tool for new employees. At the time that this case was written, no financial data were available.

The technology is available to other builders from Kenyt Technologies of India **kenyt.com**, which provides the smart real estate chatbot.

Sources: Based on Garg (2017) and **facebook.com/veragoldmark/** (accessed April 2018).

Questions for Case 12.5

1. List the benefits to VGM.
2. List the benefits to buyers.
3. What is the role of Kenyt Technologies?

Chatbots Magazine provides a three-part overview on the use of chatbots for retail and e-commerce. For details, see **chatbotsmagazine.com/chatbots-for-retail-and-e-commerce-part-three-c112a89c0b48**.

Enterprise Chatbots: Financial Services

The second area in which enterprise bots are active is financial services. Here we briefly discuss their use in banking. In Section 12.6, we present the robo financial advisors for investment.

BANKING A 2017 survey (Morgan, 2017) found that most people in the United States will bank via chatbots by 2019. Chatbots can use *predictive analytics* and *cognitive messaging* to perform tasks such as making payments. They can inform customers about personalized deals. Banks' credit cards can be advertised via chatbots on Facebook Messenger. It seems that customers prefer to deal with chatbots rather than with salespeople who can be pushy.

Examples

POSB of Singapore has an AI-driven bot on Facebook Messenger. The bot was created with the help of Kasisto, Inc. of the United States. Using actual Q&A sessions, it took IT workers 11,000 hours to create the bot. Its knowledge base was tested and verified. The bot can learn to improve its performance. Known as POSB digi-bank virtual assistant, the service is accessed via Messenger. Customers save time rather than waiting for human customer service. In the future, the service will be available on other messaging platforms. For details, see Nur (2017).

A similar application in Singapore is used by Citi Bank (by Citi Group). It can answer FAQs about people's accounts in a natural language (English). The bank is adding progressively more capabilities to its bot.

A generic banking bot is Verbal Access (from North Side Co.) that provides recommendations for banking services (see Hunt, 2017).

Enterprise Chatbots: Service Industries

Chatbots are used extensively in many services. We provide several examples in the following sections.

HEALTHCARE Chatbots are extremely active in the healthcare area, helping millions of people worldwide (Larson, 2016). Here are a few examples:

- Robot receptionists direct patients to departments in hospitals. (Similar services are available at airports, hotels, universities, government offices, and private and other public organizations.)
- Several chatbots are chatty companions for people who are elderly and sick (e.g., Zora Robotics).
- Chatbots are used in telemedicine; patients converse with doctors and healthcare professionals who are in different locations. For example, the Chinese company Baidu developed the Melody chatbot for this purpose.
- Chatbots can connect patients quickly and easily with information they need.
- Important services in the healthcare field are currently provided by IBM Watson (Chapter 6).

For more on bots for healthcare, see the end of Section 12.6.

EDUCATION Chatbot tutors are used in several countries to teach subjects ranging from English (in Korea) to mathematics (in Russia). One thing is certain: The chatbot treats all students equally. Students like the chatbots in online education as well. Machine translation of languages will enable students to take online classes in languages other than their own. Finally, chatbots can be used as private tutors.

GOVERNMENT According to Lacheca (2017), chatbots are spreading in government as a new dialog tool for use by the public. The most popular use is in providing access to government information and answering government-related questions.

TRAVEL AND HOSPITALITY Chatbots are working as tour guides in several countries (e.g., Norway). They are not only cheaper (or free) but also may know more than some human guides. Chatbots work as guides in several hotels in Japan. In hotels, they act as concierges, providing information and personalized recommendations (e.g., about restaurants). Chatbots can arrange reservations for hotel rooms, meals, and events. In busy hotels, there is frequently a wait for human concierges; chatbots are available on smartphones all the time. As with other computer services, the chatbots are fast, inexpensive, easy to reach, and always nice. They give excellent customer experience.

An example of external travel service is given in Application Case 12.6.

Chatbot Platforms

CHATBOTS INSIDE ENTERPRISES So far we have seen chatbots that are working in the external side of enterprises, mostly in customer care and marketing (e.g., the opening vignette). However, companies lately have started to use chatbots to automate tasks for supporting internal communication, collaboration, and business processes. According to

Application Case 12.6 Transavia Airlines Uses Bots for Communication and Customer Care Delivery

Background

The air travel business is very competitive, especially in Europe. There is a clear trend for younger customers to use wireless devices as well as social media sites and chatting. Customers like to communicate with travel businesses by using their preferred technology via their preferred platforms. Most popular is Facebook Messenger, where over 1.2 billion people chat, many times via their smartphones. These users today interact not only among themselves but also with the business world.

Messaging platforms such as Messenger, WhatsApp, and WeChat are becoming the norm for this customer group. Vendors are building smart apps for the messaging platforms including bots.

Transavia's Bot

Learning from other companies, Transavia decided to create a bot on Facebook Messenger. To do so, it hired the IT consultant Cognizant Digital Business unit, called Mirabean, which specializes in conversation interfaces, especially via bots. Transavia's activities business processes, marketing, and customer care were combined with Mirabean's technological experience to enable a quick deployment of the bot in weeks. It now enables real-time dialog with customers. The first application is Transavia Flight Search, which provides flight information as well as the ability to buy tickets. The system is now integrated with business processes that facilitate other transactions via the bot. Giving customers their digital tool of choice enables Transavia to increase market share and to drive growth.

Note that KLM, the owner of Transavia, was the first European airline that implemented a similar chatbot on Facebook Messenger in 2016.

Sources: Compiled from Cognizant (2017) and **transavia.com**.

Questions for Case 12.6

1. What drives consumer preference for mobile devices and chat?
2. Why was the bot placed on Facebook Messenger?
3. What were the benefits of using Cognizant?
4. What is the advantage of buying a ticket from a bot rather than from an online store?

Hunt (2017), "Enterprise and internal chatbots are revolutionizing the way companies do business." Chatbots in enterprises can do many tasks and support decision-making activities. For examples, see Newlands (2017a). Chatbots can cut costs, increase productivity, assist working groups, and foster relationships with business partners. Representative examples of chatbot tasks are:

- Help with project management.
- Handle data entry.
- Conduct scheduling.
- Streamline payments with partners.
- Advise on authorization of funds.
- Monitor work and workers.
- Analyze internal Big Data.
- Find discounted and less expensive products.
- Simplify interactions.
- Facilitate data-driven strategy.
- Use machine learning.

Facilitate and manage personal finance.

Given the large number of bots, it is not surprising that many developers started to offer tools and platforms to assist in building chatbots as discussed in Technology Insights 12.1.

TECHNOLOGY INSIGHTS 12.1 Chatbots' Platform Providers

Several companies provide platforms for building enterprise chatbots. The companies can construct chatbots fairly easily using these tools for their entry into popular messaging platforms or for their Web sites. Some of the tools have machine-learning capability to ensure that the bots learn with every interaction. According to Hunt (2017), these are some popular vendors:

1. **ChattyPeople.** This chatbot builder assists in creating bots requiring minimal programming skills. It simply allows a business to link its social media pages to its ChattyPeople account. The created bot can:
   - Arrange for payments to or from social media contacts.
   - Use major payment providers such as Apple Pay and PayPal.
   - Recognize variations in keywords.
   - Support messaging.
2. **Kudi.** This financial helper allows people to make payments to vendors directly from their messaging apps, specifically, Messenger, Skype, and Telegram and through an Internet browser. Using the bot, users can:
   - Pay bills.
   - Set bill payment reminders.
   - Transfer money by sending text messages.

   The bot is safe and it protects users' privacy. Vendors can easily install it for use.
3. **Twyla.** This chatbot building platform is for improving existing customer care and offering live chats. It acts as a messaging platform for customers who prefer to use chatting. The major objective is to free humans in HR departments from routine tasks.

The most popular platforms are:

- **IBM Watson.** This package uses a neural network of 1 billion words for excellent understanding of natural languages (e.g., English, Japanese). Watson provides free development tools, such as Java SDK, Node SDK, Pyton SDK, and iOS SDK.

- **Microsoft's Bot Framework.** Similar to IBM, Microsoft offers a variety of tools translatable into 30 languages. It is an open source. The system has three parts, Bot Connector, Developer Portal, and Bot Directory and is interconnected with Microsoft Language Understanding Intelligent Service (LUIS) that understands users' intent. The system also includes active learning technology. A simplified tool is AZURE; see Section 12.7 and Afaq (2017). For a comparative table of 25 chatbots platforms, see Davydova (2017). For a list of other platforms, see Ismail (2017).

Sources: Compiled from Hunt (2017) and Davydova (2017).

Discussion Questions

1. What is the difference between a regular enterprise bot and a platform?
2. Discuss the benefits of ChattyPeople.
3. Discuss the need for Kudi.
4. Discuss the reasons for consumers to prefer messaging platforms.

For additional information about chatbot platforms for building enterprise chatbots, see **entrepreneur.com/article/289788**.

INDUSTRY-SPECIFIC BOTS As we have seen, bots can be specialists (e.g., for investment advice, customer service) or industry-specific experts (e.g., banking, airlines). An interesting bot for the waste industry is Alto (from Bio Hi Tech Global), which enables users to communicate intelligently with industrial equipment. This helps owners of the equipment make decisions that improve performance levels, smooth maintenance routines, and facilitate communication.

Knowledge for Enterprise Chatbots

Knowledge for chatbots depends on their tasks. Most marketing and customer care bots require proprietary knowledge, which is usually generated and maintained in-house. This knowledge is similar to that of ES; in many cases, enterprise chatbots operate very similarly to ES except that the interface occurs in a natural language and frequently by voice. For example, the knowledge of Sephora's bot (opening vignette) is specific to that company and its products and is organized in a Q&A format.

On the other hand, chatbots that are used within the enterprise (e.g., to train employees or to provide advice on security or compliance with government regulations) may not be company specific. A company can buy this knowledge and modify it to fit local situations and its specific needs (as is done in ES; e.g., see Exsys Inc.). Newer chatbots use machine learning to extract knowledge from data.

PERSONAL ASSISTANTS IN THE ENTERPRISE Enterprise chatbots can also be virtual personal assistants as will be described in Section 12.5. For example, these bots can answer work-related queries and help in increasing employees' decision-making capabilities and productivity.

SECTION 12.4 REVIEW QUESTIONS

1. Describe some marketing bots.
2. What can bots do for financial services?
3. How can bots assist shoppers?
4. List some benefits of enterprise chatbots.
5. Describe the sources of knowledge for enterprise chatbots.

12.5 VIRTUAL PERSONAL ASSISTANTS

In the previous section, we introduced enterprise chatbots that can be used to conduct conversations. In marketing and sales, they can facilitate customer relationship management (CRM, execute searches for customers, provide information, and execute many specific tasks in organizations for their customers and employees. For comprehensive coverage, including research issues, see Costa et al. (2018)).

An emerging type of chatbot is designed as a virtual personal assistant for both individuals and organizations. Known as a **virtual personal assistant (VPA)**, this software agent helps people improve their work, assist in decision making, and facilitate their lifestyle. VPAs are basically extensions of intelligent software agents that interact with people. VPAs are chatbots whose major objective is to help people better perform certain tasks. At this time, millions of people are using Siri with their Apple products, Google Assistant, and Amazon's Alexa. The assistants' knowledge bases are usually universal, and they are maintained centrally in the "cloud," which makes them economical for a large number of users. Users can get assistance and advice from their virtual assistants anytime. In this section, we provide some interesting applications. The first set of applications involves virtual personal assistants, notably Amazon's Alexa and Apple's Siri and Google Assistant. O'Brien (2016) provides a discussion of what personal assistant chatbots can do for business. The second set (presented in Section 12.6) is about computer programs that act mostly as advisors on specific topics (mostly investments).

Assistant for Information Search

A major task of virtual personal assistants is to help users conduct a search by voice for information. Without the assistant, users need to surf the Internet to find information and many times abandon the search. In business situations, users can call a live customer service agent for assistance. This may be an expensive service for the vendors. Delegating the search to a machine may save sellers considerable money and make customers happy by not having to wait for the service. For example, Lenovo uses the noHold assistant in its Single Point of Search service to help customers find answers to their questions.

If You Were Mark Zuckerberg, Facebook CEO

While Siri and Alexa were in development, Zuckerberg decided to develop his own personal assistant to help him run his home and his work as the CEO of Facebook. He viewed this assistant as Jarvis from *Iron Man*. Zuckerberg trained the bot to recognize his voice and understand basic commands related to home appliances. The assistant can recognize the faces of visitors and monitor the movement of Zuckerberg's young daughter. For details, see Ulanoff (2016).

The essentials of this assistant can be seen in a 2:13 min. video at **youtube.com/watch?v=vvimBPJ3XGQ** and one (5:01 min.) at **youtube.com/watch?v=vPoT2vdVkVc**, with the narration by Morgan Freeman. Today, similar assistants are available for a minimal fee or even for free. The most well-known such assistant is Amazon's Alexa.

Amazon's Alexa and Echo

Of the several virtual personal assistants, the one considered the best in 2018 was Alexa. She was developed by Amazon to compete with Apple's Siri and is a superior product. (See Figure 12.4.) Alexa works with a smart speaker, such as Amazon's Echo (to be described later).

FIGURE 12.4 Amazon's Echo and Alexa. *Source:* McClatchy-Tribune/Tribune Content Agency LLC/ Alamy Stock Photo

Amazon's **Alexa** is a cloud-based virtual personal voice assistant that can do many things such as:

- Answer questions in several domains.
- Control smartphone operations with voice commands.
- Provide real-time weather and traffic updates.
- Control smart home appliances and other devices by using itself as a home automation hub.
- Make to-do lists.
- Arrange music in Playbox.
- Set alarms.
- Play audio books.
- Control home automation devices, as well as home appliances (e.g., a microwave).
- Analyze shopping lists.
- Control a car's devices.
- Deliver proactive notification.
- Shop for its user.
- Make phone calls and send text messages.

Alexa has the ability to recognize different voices, so it can provide personalized responses. Also, she uses a mix of speech and touch to deliver news, hail an Uber, and play games. As time passes, her capabilities and skill grow. For more capabilities, which are ever-increasing, see Johnson (2017). For what Alexa can hear and remember and how she learns, see Oremus (2018).

Watch the 3:55 min. video of how Alexa works at **youtube.com/watch?v=jCtfRdqPlbw**. For more tasks, see **cnet.com/pictures/what-can-amazon-echo-and-alexa-do-pictures/**, Mangalindan (2017), and **tomsguide.com/us/pictures-story/1012-alexa-tricks-and-easter-eggs.html**.

ALEXA'S SKILLS In addition to the standard (native) capabilities listed, people can use Alexa apps (referred to as *Skills*) to download customized capabilities to Alexa (via your smartphone). Skills are intended to teach Alexa something new.

The following are examples of Alexa's Skills (Apps):

- Call Uber and find the cost of a ride.
- Order a pizza.
- Order take-out meals.
- Obtain financial advice.
- Start a person's Hyundai Genesis car from inside her or his house (Korosec, 2016).

These skills are provided by third-party vendors; they are required to activate invocation commands. There are tens of thousands of them.

For example, a person can say, "Alexa, call Uber to pick me up at my office at 4:30 P.M." For more on Amazon's Alexa, see Kelly (2018); for its benefits, see Reisinger (2016).

Alexa is equipped with NLP user interface, so it can be activated by providing a voice command. This is done by combining the Alexa software with Amazon's intelligent speaker, Echo.

ALEXA'S VOICE INTERFACE AND SPEAKERS Amazon has a family of three speakers (or voice communication devices for Alexa: Echo, Dot, and Tag. Alexa can be accessed by a Fire TV line and some non-Amazon devices. For the relationship between Alexa and Echo, see Gikas (2016).

AMAZON'S ECHO **Echo** is a hands-free intelligent (or smart) wireless speaker that is controlled by voice. It is the hardware companion of Alexa (a software product), so the two operate hand in hand. Echo is always on, always listening. When Echo hears a question, command, or request, it sends the audio to Alexa and from there up to the cloud. Amazon's servers match responses to the questions, delivering them to Alexa as "responses to questions" in a split second. Amazon's Alexa/Echo is now available in some Ford vehicles.

Amazon Echo Dot Amazon Echo Dot is the "little brother" of Echo. It offers full Alexa functionality but has only one very small speaker. It can be linked to any existing speaker systems to provide an Echo-like experience.

Amazon Echo Tap Amazon Echo Tap is another "little brother" of Echo that can be used on the go. It is completely wireless and portable and can be charged via a charging dock.

Both Dot and Tap are less expensive than Echo, but they offer fewer functionalities and lower quality. However, people who already have good home speakers can use Dot with them. For a discussion about the three speakers, see Trusted Review at **trustedreviews.com/news/amazon-echo-show-vs-echo-2948302**.

Note: Non-Amazon speakers for Alexa are available now (e.g., Eufy Genie, from third-party vendors); some are inexpensive.

Note: Alexa was smart enough earlier to admit that she did not know an answer, but today, she will make references to third-party sources for an answer she cannot make. For details and examples, see **uk.finance.yahoo.com/news/alexa-recommend-third-party-skills-192700876.html**.

ALEXA FOR THE ENTERPRISE While the initial use of Alexa was for individual consumers, her use for business has increased. WeWork Corp. developed a platform for helping companies to integrate an Alexa skill in meeting rooms, for example. For details, see Crook (2017), and **yahoo.com/news/destiny-2-alexa-skills-let-140946575.html/**.

Apple's Siri

Siri (short for Speech Interpretation and Recognition Interface) is an intelligent virtual personal assistant and knowledge navigator. It is a part of Apple's several operating systems. It can answer questions, make recommendations, and perform some actions by delegating requests to a set of *Web services* in the "cloud." The software can adapt itself to the user's individual language, search preferences with continuing use, and return personalized results. Siri is available for free to iPhone and iPad users.

Siri can be integrated into Apple's *Siri Remote*. Using CarPlay, Siri is available in some auto brands where it can be controlled by iPhone (5 and higher). Siri 2 is the 2017–2018 model.

VIV In 2016, Dag Kittlaus, the creator of Siri, introduced Viv, "an intelligent Interface for everything." Viv is expected to be the next generation of intelligent virtual interactions (for details, see Matney, 2016). In contrast with other assistants, Viv is open to all developers (third-party ecosystem products). Viv is now a Samsung Company. In 2017, Samsung launched its own personal assistant for the Galaxy S8.

Google Assistant

Competition regarding virtual personal assistants is increasing with the improved capabilities of Google Assistant, which was developed as a competitor to Siri to fit Android smartphones. An interesting demonstration of it is available at **youtube.com/watch?v=WTMbF0qYWVs**; some advanced capabilities are illustrated in the video at **youtube.com/watch?v=17rY2ogJQQs**. For details, see Kelly (2016). The product improved dramatically in 2018 as shown in CES 2018 Conference.

Other Personal Assistants

Several other companies have virtual personal assistants. For example, Microsoft Cortana is well known. In September 2016, Microsoft combined Cortana and Bing (see Hachman, 2016). Alexa and Cortana now work together. Note that it is estimated that by the year 2022, voice-enabled personal assistants will reach 55 percent of all U.S. households. For this and the future of personal assistants, see Perez (2017).

Competition Among Large Tech Companies

Apple and Google have provided their personal assistants to hundreds of million users of their mobile devices. Microsoft has equipped over 250 million PCs with its personal assistant. Amazon's Alexa/Echo sells many more assistants than others. The competition is on voice-controlled chatbots. Their competitors view them as "the biggest thing since the iPhone."

Knowledge for Virtual Personal Assistants

As indicated earlier, the knowledge for virtual personal assistants is kept in the "cloud." The reason is that the assistants are commodities, available to millions of users, and need to provide dynamic, updated information (e.g., weather conditions, news, stock prices). When the knowledge base is centralized, its maintenance is performed in one place. This is in contrast with the knowledge of many enterprise bots, for which updating is decentralized. Thus, Siri on an iPhone will always be updated for its general knowledge by AAPL. Knowledge for the skills of Alexa has to be maintained locally or by the third-party vendors that create them.

SECTION 12.5 REVIEW QUESTIONS

1. Describe an intelligent virtual personal assistant.
2. Describe the capabilities of Amazon's Alexa.
3. Relate Amazon's Alexa to Echo.
4. Describe Echo Dot and Tap.
5. Describe Apple's Siri Google's Assistant.
6. How is the knowledge of personal assistants maintained?
7. Explain the relationship between virtual personal assistants and chatbots.

12.6 CHATBOTS AS PROFESSIONAL ADVISORS (ROBO ADVISORS)

The personal assistants described in Section 12.5 can provide much information and rudimentary advice. A special category of virtual personal assistants is designed to provide personalized professional advice in specific domains. A major area for their activities is investment and portfolio management where robo advisors operate.

Robo Financial Advisors

It is known that the vast majority of "buy" and "sell" decisions of stock trading on the major exchanges, especially by financial institutions, are made by computers. However, computers can also manage an individual's accounts in a personalized way.

According to an A. T. Kearney's survey (reported by Regan, 2015), **robo advisors** are defined as online providers that offer automated, low-cost, personalized *investment advisory* services, usually through mobile platforms. These robo advisors use algorithms that allocate, deploy, rebalance, and trade investment products. Once enrolled for the robo service, individuals enter their *investment objectives* and preferences. Then, using advanced AI algorithms, the robo will offer *alternative personalized* investments for individuals to choose from funds or exchange-traded funds [ETFs]. By conducting a dialog with the robo advisor, an AI program will refine the investment portfolio. This is all done digitally without having to talk to a live person. For details, see Keppel (2016).

Evolution of Financial Robo Advisors

The pioneering emergence of Betterment Inc. in 2010 (described later) was followed by several other companies (Future Advisor and Hedgeable in 2010 and Personal Capital, Wealthfront, and SigFig in 2011 and 2012). Other well-known companies (Schwab Intelligent Portfolios, Acorns, Vanguard RAS, and Ally) joined the crowd in 2014 and 2015. In 2016 and 2017, the brokerage houses of E*Trade and TD Ameritrade joined, as did Fidelity and Merrill Edge. There is no question that robo advisors are game-changing phenomena for the wealth management business, even though their performance so far has not been much different from that of traditional, manual, and financial services.

Robo advising companies try to cut costs by using ETFs, whose commission fees are significantly lower than that of mutual funds. Annual fees vary as does the minimum amount of required assets. Premium services are more expensive since they offer the opportunity to consult human experts (advisors 2.0), which are described next.

Robo Advisors 2.0: Adding the Human Touch

As robo advisors matured, it became clear that sometimes they could not do an effective job by themselves. Therefore, in late 2016, several of the fully automated advisors started to add what they call the *human touch* (e.g., see Eule, 2017; Huang, 2017). Companies

are adding a human option, or partner with another company. For example, UBS Wealth Management Americas has partnered with pure robo advisor SigFig.

Robo advisors with human additions vary in expertise. For example, Betterment (Plus and Permission options), Schwab Intelligent Advisory, and Vanguard Personal Advisor Service use certified financial planners (CFPs); other companies offer less expertise. For details, see Huang (2017).

Application Case 12.7 describes how Betterment has added the human touch.

QUALITY OF ADVICE PROVIDED BY ROBO ADVISORS You may wonder how good the advice from robo advisors is. The answer is that it depends on their knowledge, the type of investments involved, the inference engine of the AI machine, and so on. However, remember that the robots are not biased and are consistent. They may prove to be even

Application Case 12.7 Betterment, the Pioneer of Financial Robo Advisors

As the pioneer of financial robo advisors in 2010, Betterment created an automated platform for wealth management. Since then, it has played a leading role in a growing industry. In 2017, the company controlled more than $9 billion in assets, yielding over an 11 percent return to its 200,000 members. Like other robo advisors, Betterment appeals to investors who do not want to manage their portfolio by themselves or pay the 2 to 3 percent annual fee charged by human advisors.

The company advertises the following benefits:

- Provides unlimited professional expert advice (by the bot) anytime and anywhere.
- Provides advice from bots that contain the knowledge of human investment advisors.
- Assists investors in making decisions of how much to invest.
- Helps investors figure out how much risk to take.
- Helps in lowering investment-related tax.
- Provides actionable answers to questions.
- Advises on college savings.
- Helps plan for retirement.
- Assists in mortgage management (e.g., refinance).
- Provides personalized service via the use of investors' goal-based analysis.

Betterment has no account minimum (competitors require up to $100,000).

Each investor's portfolio is automatically adjusted to market conditions to meet his or her goals. All portfolios are built and managed by AI algorithms.

Premium Service—Adding the Human Touch

Like **Amazon.com** and Expedia, which started as pure online companies and later added physical commerce, in 2017 Betterment added what it calls a human touch; its Plus service is offered to customers with assets of over $100,000 who are willing to pay an annual fee of 0.4 percent for this service. Using it, customers can interact with human advisors in addition to the automated bot. An even better service is the company's Premium level, which requires $250,000 in assets and charges 0.5 percent in fees.

While the quality of the automated service is getting better with added knowledge (machine learning), complex situations that require human intervention still remain. This is where the Plus and Premium services enter the picture. Several competitors also have added the human touch to their offering.

Sources: Compiled from O'Shea (2017), Eule (2017), and **betterment.com** (accessed April 2018).

Questions for Case 12.7

1. What are Betterment's benefits to investors?
2. Compare Betterment to its major competitors (see Eule, 2017).
3. What are the benefits of adding the human touch (i.e., compared to pure automation and only human service)?
4. Find some new information about Betterment. Write a report.

better than humans at one of the most important aspects in investment advising: know how to legally minimize the related tax. This implies that institutional-grade tax-loss harvesting is now within the reach of all investors. By contrast, some people believe that it is difficult to replace investment brokers with robots. De Aenlle (2018) believes that humans are still dominating advisory services (see the example of Nordea Bank by Pohjanpalo, 2017).

For a list of the best robo advisors, see Eule (2017), O'Shea (2016), and **investorjunkie.com/35919/roboadvisors**. For comprehensive coverage of robo advisors in finance and investment, including the major companies in the advisory industry, see McClellan (2016).

An emerging commercial robo advisor is being developed at Cornell University under the name Gsphere. In addition, robo advisors appear in countries other than the United States (e.g., Marvelstone Capital in Singapore).

FINANCIAL INSTITUTIONS AND THEIR COMPETITION Several large financial institutions and banks have reacted to robo advisors by creating their own or partnering with them. It is difficult to assess the winners and losers in this competition because there are no sufficient long-term data. So far it seems that customers like robo advisors, basically because they cost as little as 10 percent of full-service human advisors. For a discussion and data, see Marino (2016). Note that some observers point to the danger of using robo advisors in a declining stock market due to their use of ETFs.

Managing Mutual Funds Using AI

Many institutions and some individual investors buy stocks using AI algorithms. Some people prefer to buy a mutual fund that picks its holding with AI. EquBot is such a fund (its symbol is AIEQ). Its 2017 performance was above average.

The AI algorithms used by EquBot can process 1 million pieces of data each day. They follow 6,000 companies. For details, see Ell (2018).

Other Professional Advisors

In addition to investment advisors, there are several other types of robo advisors ranging from travel to medicine to legal areas.

The following are examples of noninvestment advisors:

- **Computer operations.** To cut costs, major computer vendors (hardware and software) try to provide users with self-guides to solve encountered problems. If users cannot get help from the guides, they can contact live customer service agents. This service may not be available in real time, which can upset customers. Live agents are expensive, especially when provided **24/7**. Therefore, companies are using interactive virtual advisors (or assistants).

  As an example, Lenovo Computers use a generic bot called noHold's AI to provide assistance to customers as a single point of help for conducting a search.
- **Travel.** Several companies provide advice on planning future national and international trips.

  For example, Utrip (**utrip.com**) helps plan European trips. Based on their stated objectives, travelers get recommendations for what to visit in certain destinations. The service is different from others in that it customizes trips.
- **Medical and health advisors.** A large number of health and medical care advisors operate in many countries. An example is Ad a Health of Germany. Founded in late 2017 as a chatbot, it assists people in activities such as deciphering their ailments and can connect patients to live physicians. This can be the future of health in adding bot-based patient-doctor collaboration.

A list of the top useful chatbots as of 2017 is provided by TalKing (2017). It includes:

- *Health Tap* acts like a medical doctor by providing a solution to common symptoms provided by patients.
- *YourMd* is similar to Health Tap.
- *Florence* is a personal nurse available on Facebook Messenger.

  Other bots include OneStopHealth, HealthBot, GYANT, Buoy, Bouylon, and Mewhat.
- Bots are acting as companions (e.g., Endurance for dementia patients). In Japan, bots that look and feel like dogs are very popular companions for elderly people. Several bots are designed to increase patient engagement. For example, Lovett (2018) reports that a bot for patient engagement increased patients' response rate to a flu shot campaign by 30 percent. Finally, the classic pioneering bot, ELIZA, acted as a very naïve psychologist.
- **Shopping advisors (shopbots).** Shopbots can act as shopping advisors. An example is Shop Advisor (see **shopadvisor.com/our-platform**. It is a comprehensive platform that includes three components to help companies attract customers. The platform is a self-learning system that improves its operation over time. Its components are:

  1. Product intelligence, which processes complex and diverse product data. It includes a competitive analysis.
  2. Context intelligence, which collects and catalogs contextual data points about marketing facilities and inventories in different locations.
  3. Shopper intelligence, which studies consumers' actions related to different magazines, mobile apps, and Web sites.

There are thousands of other shopping advisors. Sephora (opening vignette) has several of them. There are chatbots for Mercedes cars and for top department stores such as Nordstrom, Saks, and DFS. The use of shopping chatbots is increasing rapidly due to the use of mobile shopping and mobile chatting on social networks. Marketers, as we stated earlier, can collect customer data and deliver targeted ads and customer service to specific customers.

Another trend that facilitates online shopping with the assistance of bots is the increase in the number of virtual personal shopping assistants. Users only have to tell Alexa by voice, for example, to buy something for them. Better than that, they can use their smartphones from anywhere to tell Alexa to go shopping. Ordering via voice directly from vendors (e.g., delivery of pizzas) is becoming popular. In addition to chatbots that operate by sellers, there are bots for providing advice on what and where to buy.

Example: Smart Assistant Shopping Bots

Shopping bots ask a few questions to understand what a customer needs and prefers. Then they recommend the best match for the customer. This makes customers feel they are receiving personalized service. The assistance simplifies the customer's decision-making process. Smart assistants also offer advice on issues of concern to customers via Q&A conversations. For a guided test, go to a demo at **smartassistant.com/advicebots**. Note that these bots are essentially recommendation systems and that users need to ask for advice whereas other recommendation systems (e.g., that of **Amazon.com**) provide advice even when users do not ask for it.

A well-known global shopping assistant in the area of fashion is Alibaba's Fashion AI. It helps customers who shop in stores. When shoppers enter a fitting room, the AI Fashion Consultant goes into action. For details of how this is done, see Sun (2017).

Another type of shopping advisor works as a virtual personal advisor to shoppers. This type was developed from traditional e-commerce intelligent agents, such as **bizrate.com** and **pricegrabber.com**.

IBM Watson

Probably the most knowledgeable virtual advisor is IBM Watson (see Chapter 6). Some examples of its use follow:

- Macy's developed a service, Macy's On Call, to help customers navigate its physical stores while they shop. Using location-based software, the app knows where they are in the store. By using smartphones, customers can ask questions regarding products and services in the stores and then receive a customized response from the chatbot.
- Watson can help physicians make a diagnosis (or verify one) quickly and suggest the best treatment. Watson's Medical Advisor can analyze images very fast and look for things that physicians may miss. Watson already is used extensively in India where there is a large shortage of doctors.
- Deep Thunder provides accurate weather-forecasting service.
- Hilton Hotels are using Watson-based "Connie Robot" in their front desks. Connie did a superb job in experiments, and its service is improving.

Clark (2016) reports that 1 billion people will use Watson by 2018. This is in part because IBM Watson is coming to smartphones as an advisor. For more, see Noyes (2016).

SECTION 12.6 REVIEW QUESTIONS

1. Define *robo advisor*.
2. Explain how robo advisors work for investments.
3. Discuss some of the shortcomings of robo advisors for investments.
4. Explain the people-machine collaboration in robo advising.
5. Describe IBM Watson as an advisor.

12.7 IMPLEMENTATION ISSUES

Several implementation issues are unique to chatbots and personal assistants. Examples of representative systems are described next.

Technology Issues

Many chatbots, including virtual personal assistants, have imperfect (but improving) voice recognition. There is no good feedback system yet for voice recognition systems to tell users, in real time, how well it understands them. In addition, voice recognition systems may not know when to do a current task and need to ask for human intervention.

Chatbots that are internal to organizations need to be connected to an NLP system. This may be a problem, but a bigger one may exist when chatbots are connected to the Internet, due to security and connectivity difficulties.

Some chatbots need to be multilingual. Therefore, they need to be connected to a machine language translator.

Disadvantages and Limitations of Bots

The following are points (which were observed at the time this book was written during 2017 and 2018) regarding bots' disadvantages and limitations; some will disappear with time:

- Some bots provide inferior performance, at least during their initiation, making users frustrated.
- Some bots do not properly represent their brand. Poor design may result in poor representation.
- The quality of AI-based bots depends on the use of complex algorithms that are expensive to build and use.
- Some bots are not convenient to use.
- Some bots operate in an inconsistent manner.
- Enterprise chatbots pose great security and integration challenges.

For methods to eliminate some of the disadvantages and limitations, see Kaya 2017.

VIRTUAL ASSISTANTS UNDER ATTACK Cortana, Siri, Alexa, and Google Assistant are under attack by people who are enraged at machines in general, or just like to make fun of them. In some cases, the bots' administrators try to compose a response to the attacks; in other cases, some machines provide senseless responses to the senseless attacks.

Quality of Chatbots

While the quality of most systems is not perfect, it is improving over time. However, the quality of those that retrieve information for users and are properly programmed can do a perfect job. Generally speaking, the more a company invests in acquiring or leasing a chatbot, the better its accuracy will be. In addition, bots that serve a large number of people, such as Alexa and Google Assistant, exhibit an increasing level of accuracy.

QUALITY OF ROBO ADVISORS Given the short time since the emergence of robo advisors for financial services, it is difficult to assess the quality of their advice. Backend Benchmarking publishes a quarterly report (**theroboreport.com**) regarding robo advisor companies. Some reports are free. According to this service, Schwab's Intelligent Portfolio Robot was the top performer in 2017. However, note that portfolio performance needs to be measured for the long run (e.g., 5 to 10 years).

A major issue when engaging bots is the potential loss of human touch. It is needed to build trust and answer complex questions so customers can understand bots' answers. Also, bots cannot bring empathy or a sense of friendship. According to Knight (2017b), there is a solution to this. First, bots should perform only tasks that they are suited to do. Second, they should provide a visible benefit to the customer. Finally, because the bots face customers, the interactions must be fully planned to make sure the customers are happy.

In addition, note that robo advisors provide personalized advice. For information as to which robo may be best for you based on your objectives, see Eule (2017), who also provides a scorecard for the leading companies in the field. Finally, Gilani (2016) provides a guide for robo advisors as well as their possible dangers.

MICROSOFT'S TAY Tay was a Twitter-based chatbot that failed and was discontinued by Microsoft. It collected information from the Internet, but Microsoft had not given the bot the knowledge of how to deal with some inappropriate material used on the Internet (e.g., trolls, fake news). Therefore, Tay's output was useless and frequently offended its users. As a result, Microsoft discontinued the service of Tay.

Setting Up Alexa's Smart Home System

Alexa is useful in controlling smart homes. Crist (2017) proposed a six-step process for how to use Alexa in smart homes:

1. Get a speaker (e.g., Echo).
2. Think about the location of the speaker.
3. Set up the smart home devices.
4. Sync related gadgets with Alexa.
5. Set up group and scene.
6. Fine-tune during the process.

These steps are demonstrated at **cnet.com/uk/how-to/how-to-get-started-with-an-alexa-smart-home/**.

Constructing Bots

Earlier, we presented some companies that provide development platforms for chatbots. In addition, several companies can build bots for users, so they can also build a simple bot by themselves. A step-by-step guide with the tools used is provided by Ignat (2017). The bot was constructed on Facebook Messenger. Another guide for creating a Facebook Messenger bot is provided by Newlands (2017b), who suggested the following steps:

1. Give it a unique name.
2. Give customers guides on how to build a bot and how to converse with it.
3. Experiment in making a natural conversation flow.
4. Make the bot sound smart, but use simple terminology.
5. Do not deploy all features at the same time.
6. Optimize and maintain the bot to constantly improve its performance.

There are several free sources for building chatbots. Most of them include "how-to" instructions. Several messaging services (e.g., Facebook Messenger, Telegraph) provide both chatbot platforms as well as their own chatbots. For a 2017 list of enterprise chatbot platforms and their capabilities, see **entrepreneur.com/article/296504**.

USING MICROSOFT'S AZURE BOT SERVICE Azure is a comprehensive but not a very complex bot builder. Its Bot Service provides five templates for quick and easy creation of bots. According to **docs.microsoft.com/en-us/bot-framework/azure-bot-service-overview/**, any of the templates shown in Table 12.1 can be used.

For a detailed tutorial for creating bots, see "Create a Bot with Azure Bot Service" at **docs.microsoft.com/en-us/bot-framework/azure-bot-service-overview/**.

TABLE 12.1 Azure's Templates

Template	Description
Basic	Creates a bot that uses dialogues to respond to user input.
Form	Creates a bot that collects input from users via a guided conversation that is created using Form Flow.
Language understanding	Creates a bot that uses natural language models (LUIS) to understand user intent.
Proactive	Creates a bot that uses Azure Functions to alert users of events.
Question & Answer	Creates a bot that uses a knowledge base to answer users' questions.

Note: Microsoft also provides a bot framework on which bots can be constructed (similar to that of Facebook Messenger). For Microsoft's Bot and a tutorial, see Afaq (2017).

Chapter Highlights

- Chatbots can save organizations money, provide a 24/7 link with customers and/or business partners, and are consistent in what they say.
- An expert system was the first commercially applied AI product.
- ES transfer knowledge from experts to machines so the machines can have the expertise needed for problem solving.
- Classical ES use business rules to represent knowledge and generate answers to users' questions from it.
- The major components of ES are knowledge acquisition, knowledge representation, knowledge base, user interface, and interface engine. Additional components may include an explanation subsystem and a knowledge-refining system.
- ES help retain scarce knowledge in organizations.
- New types of knowledge systems are superior to classical ES, making ES disappear.
- We distinguish three major types of chatbots: enterprise, virtual personal assistants, and robo advisors.
- A relatively new application of knowledge systems is the virtual personal assistant. Major examples of such assistants are Amazon's Alexa, Apple's Siri, and Google's Assistant.
- Knowledge for virtual personal assistants is centrally maintained in the "cloud" and it is usually disseminated via a Q&A dialog.
- Personal assistants can receive voice commands that they can execute.
- Personal assistants can provide personalized advice to their owners.
- Special breeds of assistants are personal advisors, such as robo advisors, that provide personalized advice to investors.
- Recommenders today use several AI technologies to provide personalized recommendations about products and services.
- People can communicate with chatbots via written messages, voice, and images.
- Chatbots contain a knowledge base and a natural language interface.
- Chatbots are used primarily for information search, communication and collaboration, and rendering advice in limited, specific domains.
- Chatbots can facilitate online shopping by providing information and customer service.
- Chatbots work very well with messaging systems (e.g., Facebook Messenger, WeChat).
- Enterprise chatbots serve customers of all types and can work with business partners. They can also serve organizational employees.
- Virtual personal assistants (VPAs) are designed to work with individuals and can be customized for them.
- VPAs are created as "native" products for the masses.
- A well-known VPA is Amazon's Alexa that is accessed via a smart speaker called Echo (or other smart speakers).
- VPAs are available from several vendors. Well known are Amazon's Alexa, Apple's Siri, and Google's Assistant.
- VPAs can specialize in specific domains and work as investment advisors.
- Robo advisors provide personalized online investment advice at a much lower cost than human advisors. So far, the quality seems to be comparable.
- Robo advisors can be combined with human advisors to handle special cases.

Key Terms

Alexa
chatbot
Echo
expert systems
Google's Assistant
recommendation systems
robo advisors
Siri
virtual personal assistant (VPA)

Questions for Discussion

1. Some people say that chatbots are inferior for chatting. Others disagree. Discuss.
2. Discuss the financial benefits of chatbots.
3. Discuss how IBM Watson will reach 1 billion people by 2018 and what the implications of that are.
4. Discuss the limitation of chatbots and how to overcome them.
5. Discuss what made ES popular for almost 30 years before their decline.
6. Summarize the difficulties in knowledge acquisition from experts (also consult Chapter 2).
7. Compare the ES knowledge-refining system with knowledge improvement in machine learning.
8. Discuss the difference of enterprises' use of chatbots internally and externally.
9. Some people say that without a virtual personal assistant, a home cannot be smart. Why?
10. Compare Facebook Messenger virtual assistant project M with that of competitors.
11. Examine Alexa's skill in ordering drinks from Starbucks.
12. Discuss the advantages of robo advisors over human advisors. What are the disadvantages?
13. Explain how marketers can reach more customers with bots.
14. Are robo advisors the future of finance? Debate; start with Demmissie (2017).
15. Research the potential impact of chatbots on work and write a summary.

Exercises

1. Compare the chatbots of Facebook and WeChat. Which has more functionalities?
2. Enter **nuance.com** and find information about Dragon Medical Advisor. Describe its benefits. Write a report.
3. Enter **shopadvisor.com/our-platform** and review the platform's components. Examine the product's capabilities and compare them with those of two other shopping advisors.
4. Enter **chatbots.org/** and join a forum of your interest. Also explore research issues of your interest. Write a report.
5. There is intense competition between all major tech companies regarding their virtual personal assistants. New innovations and capabilities appear daily. Research the status of these assistants for Amazon, Apple, Microsoft, Google, and Samsung. Write a report.
6. Some people believe that chatbots will change how people interact with the Internet and browse online. Prepare a report regarding this.
7. Explain why is Amazon's Echo needed to work with Alexa? Read **howtogeek.com/253719/do-i-need-an-amazon-echo-to-use-alexa/**. Write a report.
8. Find out how Simon Property Group is using chatbots across over 200 shopping malls. Write about the benefits to different types of users and to the company.
9. Read recent information about enterprise bots. Write a report.
10. Enter **gravityinvestments.com/digital-advice-platform-demo**. Would you invest in this project? Research and write a report.
11. Enter visirule.co.uk and find all products it has for expert systems. List them and write a short report.
12. Research the role of chatbots in helping patients with dementia.
13. Find information on the Baidu's Melody chatbot and how it works with Baidu Doctor.
14. Pose a question related to a chatbot on **quora.com**. Summarize the answers received in a report.
15. Nina is an intelligent chatbot from Nuance Communication Inc. that works for Alexa Internet of Things (IoT), smart homes, and more. Find information and write a report about Nina's capabilities and benefits.
16. Microsoft partners with the government of Singapore to develop chatbots for e-services. Find out how this is done.
17. Study the Tommy Hilfiger Facebook Messenger bot. Find out how it is (and was) used in the company's marketing campaigns.
18. Two comprehensive building tools for chatbots are Botsify and Personality Forge (**personalityforge.com**). Compare the tools. Write a report.
19. Find information about the Alibaba-backed robo advisor Youyu by Yunfeng's Investment. What is unique about this service? Start by visiting **http://www.international-adviser.com/news/1035281/alibaba-backed-retail-robo-adviser-youyu-launches-honk-kong/**.
20. Enter **exsys.com**. Select three case studies and explain why they were successful.
21. It is time now to build your own bot. Consult with your instructor about which software to use. Have several bots constructed in your class and compare their capabilities. Use Microsoft's Azure if you have some programming experience.

References

Afaq, O. "Developing a Chatbot Using Microsoft's Bot Framework, LUIS and Node.js (Part 1)." *Smashing Magazine*, May 30, 2017. **smashingmagazine.com/2017/05/chatbot-microsoft-bot-framework-luis-nodejs-part1/** (accessed April 2018).

Aggarwal, C. *Recommended Systems: The Textbook*. [eTextbook]. New York, NY: Springer, 2016.

Arora, S. "Recommendation Engines: How Amazon and Netflix Are Winning the Personalization Battle." *Martech Advisor*, June 28, 2016.

Arthur, R. "Sephora Launches Chatbot on Messaging App Kik." *Forbes*, March 30, 2016.

Bae, J. "Development and Application of a Web-Based Expert System Using Artificial Intelligence for Management of Mental Health by Korean Emigrants." *Journal of Korean Academy of Nursing*, April 2013.

Beaver, L. "Chatbots Explained: Why Businesses Should Be Paying Attention to the Chatbot Revolution." *Business Insider*, March 4, 2016.

CBS News. "LinkedIn Adding New Training Features, News Feeds and 'Bots.'" *CBS News*, September 22, 2016. **cbsnews.com/news/linkedin-adding-new-training-features-news-feeds-and-bots** (accessed April 2018).

Clark, D. "IBM: A Billion People to Use Watson by 2018." *The Wall Street Journal*, October 26, 2016.

Cognizant. "Bot Brings Transavia Airlines Closer to Customers." *Cognizant Services*, 2017. **https://www.cognizant.com/content/dam/Cognizant_Dotcom/landing-page-resources/transavia-case-study.pdf** (accessed April 2018).

Costa, A., et al. (eds.) *Personal Assistants: Emerging Computational Technologies (Intelligent Systems Reference Library)*. New York, NY: Springer, 2018.

Crist, R. "How to Get Started with an Alexa Smart Home." *CNET*, July 5, 2017. **cnet.com/how-to/how-to-get-started-with-an-alexa-smart-home/** (accessed April 2018).

Crook, J. "WeWork Has Big Plans for Alexa for Business." *TechCrunch*, November 30, 2017.

Davydova, O. "25 Chatbot Platforms: A Comparative Table." *Chatbots Journal*, May 11, 2017.

De Aenlle, C. "A.I. Has Arrived in Investing. Humans Are Still Dominating." *The New York Times*, January 12, 2018.

Demmissie, L. "Robo Advisors: The Future of Finance." *The Ticker Tape*, March 13, 2017.

Ell, K. "ETFs Powered by Artificial Intelligence Are Getting Smarter, Says Fund Co-Founder." *CNBC News*, January 23, 2018.

Eule, A. "Rating the Robo-Advisors." *Barron's*, July 29, 2017.

Ferron, E. "Mobile 101: What Are Bots, Chatbots and Virtual Assistants?" *New Atlas*, February 16, 2017. **newatlas.com/what-is-bot-chatbot-guide/47965/** (accessed April 2018).

Garg, N. "Case Study: How Kenyt Real Estate Chatbot Is Generating Leads." *Medium*, June 22, 2017.

Gikas, M. "What the Amazon Echo and Alexa Do Best." *Consumer Reports*, July 29, 2016. **consumerreports.org/wireless-speakers/what-amazon-echo-and-alexa-do-best** (accessed April 2018).

Gilani, S. "Your Perfectly Diversified Portfolio Could Be in Danger—Here's Why." *Money Morning @ Wall Street*, December 6, 2016.

Griffiths, T. "Using Chatbots to Improve CRM Data: A WeChat Case Study." *Half a World*, November 16, 2016.

Guynn, J. "Zuckerberg's Facebook Messenger Launches 'Chat Bots' Platform." *USA Today*, April 12, 2016.

Hachman, M. "Microsoft Combines Cortana and Bing with Microsoft Research to Accelerate New Features." *PCWorld*, September 29, 2016.

Huang, N. "Robo Advisers Get the Human Touch." *Kiplinger's Personal Finance*, September 2017.

Hunt, M. "Enterprise Chatbots and the Conversational Commerce Revolutionizing Business." *Entrepreneur*, July 3, 2017.

Ignat, A. "Iggy—A Chatbot UX Case Study." *Chatbot's Life*, August 9, 2017. **chatbotslife.com/iggy-a-chatbot-ux-case-study-b5ac0379029c/** (accessed April 2018).

Ismail, K. "Top 14 Chatbot Building Platforms of 2014." *CMS Wire*, December 19, 2017.

Johnson, K. "Everything Amazon's Alexa Learned to Do in 2017." *Venturebeat.com*, December 29, 2017.

Kaya, E. *Bot Business 101: How to Start, Run & Grow Bot/AI Business*. Kindle Edition. Seattle, WA: Amazon Digital Services, 2017.

Kelly, H. "Amazon wants Alexa everywhere." *CNN Tech*, September 22, 2018.

Kelly, H. "Battle of the Smart Speakers: Google Home vs. Amazon Echo." *CNN Tech*, May 20, 2016. **money.cnn.com/2016/05/20/technology/google-home-amazon-echo/index.html?iid=EL** (accessed April 2018).

Keppel, D. *Best Robo-Advisor: Ultimate Automatic Wealth Management*. North Charleston, SC: Create Space Pub., 2016.

Knight, K. "Expert: Bots May Be a Marketers New Best Friend." *BizReport*, December 7, 2017a.

Knight, K. "Expert: How to Engage Chatbots Without Losing the Human Touch." *BizReport*, February 13, 2017b.

Knight, K. "Report: Over Half of Millennials Have or Will Use Bots." *Biz Report*, February 24, 2017c.

Korosec, K. "Start Your Car from Inside Your Home Using Amazon's Alexa." *Fortune.com*, August 18, 2016.

Lacheca, D. "Conversational AI Creates New Dialogues for Government." *eGovInnovation*, October 24, 2017.

Larson, S. "Baidu Is Bringing AI Chatbots to Healthcare." *CNNTech*, October 11, 2016.

Lovett, L. "Chatbot Campaign for Flu Shots Bolsters Patient Response Rate by 30%." *Healthcareitnews.com*, January 24, 2018.

Mah, P. "The State of Chatbots in Marketing." *CMOInnovation*, November 4, 2016.

Makadia, M. "Benefits for Recommendation Engines to the Ecommerce Sector." *Business 2 Community*, January 7, 2018.

Mangalindan, J. P. "RBC: Amazon Has a Potential Mega-Hit on Its Hand." *Yahoo! Finance*, April 25, 2017.

Marino, J. "Big Banks Are Fighting Robo-Advisors Head On." *CNBC News*, June 26, 2016.

Matney, L. "Siri-Creator Shows Off First Public Demo of Viv, 'The Intelligent Interface for Everything.'" *Tech Crunch*, May 9, 2016. **techcrunch.com/2016/05/09/siri-creator-shows-off-first-public-demo-of-viv-the-intelligent-interface-for-everything** (accessed April 2018).

McClellan, J. "What the Evolving Robo Advisory Industry Offers." *AAII Journal*, October 2016.

Morgan, B. "How Chatbots Improve Customer Experience in Every Industry: An Infograph." *Forbes*, June 8, 2017.

Newlands, M. "How to Create a Facebook Messenger Chatbot for Free Without Coding," *Entrepreneur*, March 14, 2017a.

Newlands, M. "10 Ways Enterprise Chatbots Empower CEOs." *MSN.com*, August 9, 2017b. **msn.com/en-us/money/smallbusiness/10-ways-enterprise-chatbots-empower-ceos/ar-AApMgU8** (accessed April 2018).

Noyes, K. "Watson's the Name, Data's the Game." *PCWorld*, October 7, 2016.

Nur, N. "Singapore's POSB Launches AI-Driven Chatbot on Facebook Messenger." *MIS Asia*, January 19, 2017.

O'Brien, M. "What Can Chatbots Do for Ecommerce?" *ClickZ.com*, April 11, 2016.

Oremus, W. "When Will Alexa Know Everything?" *Slate.com*, April 6, 2018.

O'Shea, A. "Best Robo-Advisors: 2016 Top Picks." *NerdWallet*, March 14, 2016.

O'Shea, A. "Betterment Review 2017." *NerdWallet*, January 31, 2017.

Perez, S. "Voice-Enabled Smart Speakers to Reach 55% of U.S. Households by 2022, Says Report." *Tech Crunch*, November 8, 2017.

Pohjanpalo, K. "Investment Bankers Are Hard to Replace with Robots, Nordea Says." *Bloomberg*, November 27, 2017.

Popper, B. "How Netflix Completely Revamped Recommendations for Its New Global Audience." *The Verge*, February 17, 2016. **theverge.com/2016/2/17/11030200/netflix-new-recommendation-system-global-regional** (accessed April 2018).

Quoc, M. "10 Ecommerce Brands Succeeding with Chatbots." *A Better Lemonade Stand*, October 23, 2017.

Radu, M. "How to Pay Less for Advertising? Use Baro—An Ad Robot for Campaigns Optimization." *150sec.com*, August 18, 2016.

Rayome, A. "How Sephora Is Leveraging AR and AI to Transform Retail and Help Customers Buy Cosmetics." *TechRepublic*, February, 2018, "If This Model Is Right." *Bloomberg Business*, June 18, 2015.

Reisinger, D. "10 Reasons to Buy the Amazon Echo Virtual Personal Assistant." Slide Show. *eWeek*, February 9, 2016.

Schlicht, M. "The Complete Beginner's Guide to Chatbots." *Chatbots Magazine*, April 20, 2016. **chatbotsmagazine.com/the-complete-beginner-s-guide-to-chatbots-8280b7b906ca** (accessed April 2018).

StartUp. "How Netflix Uses Big Data." *Medium.com*, January 12, 2018. **medium.com/swlh/how-netflix-uses-big-data-20b5419c1edf** (accessed April 2018).

Sun, Y. "Alibaba's AI Fashion Consultant Helps Achieve Record-Setting Sales." *MITTechnology Review*, November 13, 2017.

TalKing. "Top Useful Chatbots for Health." *Chatbots Magazine*, February 7, 2017.

Taylor, S. "Very Human Lessons from Three Brands That Use Chatbots to Talk to Customers." *Fast Company*, October 21, 2016. **fastcompany.com/3064845/human-lessons-from-brands-using-chatbots** (accessed April 2018).

Ulanoff, L. "Mark Zuckerberg's AI Is Already Making Him Toast." *Mashable*, July 22, 2016.

CHAPTER 13

The Internet of Things as a Platform for Intelligent Applications

LEARNING OBJECTIVES

- Describe the IoT and its characteristics
- Discuss the benefits and drivers of IoT
- Understand how IoT works
- Describe sensors and explain their role in IoT applications
- Describe typical IoT applications in a diversity of fields
- Describe smart appliances and homes
- Understand the concept of smart cities, their content, and their benefits
- Describe the landscape of autonomous vehicles
- Discuss the major issues of IoT implementation

The Internet of Things (IoT) has been in the technology spotlight since 2014. Its applications are emerging rapidly across many fields in industry, services, government, and the military (Manyika et al., 2015). It is estimated that 20 to 50 billion "things" will be connected to the Internet by 2020–2025. The IoT connects large numbers of smart things and collects data that are processed by analytics and other intelligent systems. The technology is frequently combined with artificial intelligence (AI) tools for creating smart applications, notably autonomous cars, smart homes, and smart cities.

13.1 OPENING VIGNETTE: CNH Industrial Uses the Internet of Things to Excel

CNH Industrial N.V. (CNH) is a Netherlands-based global manufacturer of vehicles for agriculture, construction, and commercial markets. The company produces and services more than 300 types of vehicles and operates in 190 countries where it employs over 65,000 people. The company's business is continuously growing while operating in a very competitive environment.

THE PROBLEM

To manage and coordinate such a complex business from its corporate office in London, the company needed a superb communication system as well as effective analytical capabilities and a customer service network. For example, the availability of repair parts is critical. Customers' equipment does not work until a broken part is replaced. Competitive pressures are very strong, especially in the agriculture sector where weather conditions, seasonality, and harvesting pressure may complicate operations. Monitoring and controlling equipment properly is an important competitive factor. Predicting equipment failures is very desirable. Rapid connectivity with customers and the equipment they purchase from CNH is essential as are efficient data monitoring and data collection. Both CNH and its customers need to make continuous decisions for which real-time flow of information and communication is essential.

THE SOLUTION

Using PTC Transformational Inc. as an IoT, vendor, CNH implemented an IoT-based system with internal structural transformation in order to solve its problems and reshape its connected industrial vehicles. The initial implementation was in the agricultural sector. The details of the implementation are provided by PTC, Inc. (2015). The highlights of this IoT are summarized next.

- Connects all vehicles (those that are equipped with sensors and are connected to the system) in hundreds of locations worldwide to CNH's command and control center. This connection enables monitoring performance.
- Monitors the products' condition and operation as well as their surrounding environments through sensors. It also collects external data, such as weather conditions.
- Enables customization of products' performance at customers' sites.
- Provides the data necessary for optimizing the equipment's operation.
- Analyzes the performance of the people who drive CNH's manufactured vehicles and recommends changes that can improve the vehicles' efficiency.
- Predicts the range of the fuel supply in the vehicles.
- Alerts owners to the needs (and timing) of preventive maintenance (e.g., by monitoring usage and/or predicting failures) and orders the necessary parts for such service. This enables proactive and preventive maintenance practices.
- Finds when trucks are overloaded (too much weight), violating CNH's warranty.
- Provides fast diagnosis of products' failures.
- Enables the delivery of trucks on schedule by connecting them to planners and with delivery sources and destinations.
- Helps farmers to optimally plan the entire farming cycle from preparing the soil to harvesting (by analyzing the weather conditions).
- Analyzes collected data and compares them to standards.

All of this is done mostly wirelessly.

THE RESULTS

According to Marcus (2015), CNH *halved the downtime* of its participating equipment at customer sites by using the IoT. Parts for incoming orders can be shipped very quickly. IoT use also helped farmers monitor their fields and equipment to improve efficiency. The company is now showing customers less effective examples of operations and superb operating practices. In addition, product development benefits from the analysis of collected data.

Sources: Compiled from PTC, Inc. (2015), Marcus (2015), and **cnhindustrial.com/en-us/pages/homepage.aspx**.

QUESTIONS FOR THE OPENING VIGNETTE

1. Why is the IoT the only viable solution to CNH's problems?
2. List and discuss the major benefits of IoT.
3. How can CNH's product development benefit from the collected data about usage?
4. It is said that the IoT enables telematics and connected vehicles. Explain.
5. Why is IoT considered the "core of the future business strategy"?
6. It is said that the IoT will enable new services for CNH (e.g., for sales and collaboration with partners). Elaborate.
7. View Figure 13.1 (The process of IoT) and relate it to the use of IoT at CNH.
8. Identify decision support possibilities.
9. Which decisions made by the company and its customers are supported by IoT?

WHAT WE CAN LEARN FROM THIS VIGNETTE

First, we learned how IoT provides an infrastructure for new types of applications that connect thousands of items to a decision-making center.

Second, we learned about the flow of data collected by sensors from vehicles and the environment around them and their transmittal for analytical processing.

Third, the manufacturer of the vehicles and their owners and users can enjoy tremendous benefits from using the system.

Finally, this, IoT provides an efficient communication and collaboration framework for decision makers, the manufacturer's organization, and the users of the purchased equipment.

In this chapter, we elaborate on the technologies involved and the process of the IoT operation. We also describe its major application in enterprises, homes, smart cities, and autonomous (smart) vehicles.

13.2 ESSENTIALS OF IoT

The **Internet of Things (IoT)** is an evolving term with several definitions. In general, IoT refers to a computerized network that connects many objects (people, animals, devices, sensors, buildings, items) each with an embedded microprocessor. The objects are connected, mostly wirelessly, to the Internet forming the IoT. The IoT can exchange data and allow communication among the objects and with their environments. That is, the IoT allows people and things to be interconnected anytime and anyplace. Embedded sensors that collect and exchange data make up a major portion of the objects and the IoT. That is, IoT uses *ubiquitous computing*. Analysts predict that by the year 2025, more than 50 billion devices (objects) will be connected to the Internet, creating the backbone

of IoT applications. The challenges and opportunities of this disruptive technology (e.g., for cutting costs, creating new business models, improving quality) are discussed in an interview with Peter Utzschneider, vice president of product management for Java at Oracle (see Kvitka, 2014). In addition, you can join the conversations at **iotcommunity.com**. For Intel's vision of a fully connected world, see Murray (2016).

Embedding computers and other devices that can be switched on and off into active items anywhere and connecting all devices to the Internet (and/or to each other) permit extensive communication and collaboration between users and items. By connecting many devices that can talk to each other, one can create applications with new functionalities, increase the productivity of existing systems, and drive the benefits discussed later. This kind of interaction opens the door to many applications. For business applications of the Internet of Things, see Jamthe (2016). In addition, check the "Internet of Things Consortium" (iofthings.org) and its annual conferences. For an infographic and a guide, see **intel.com/content/www/us/en/internet-of-things/infographics/guide-to-iot.html**.

Definitions and Characteristics

There are several definitions of IoT.

Kevin Ashton, who is credited with the term the "Internet of Things," provided the following definition: "The Internet of Things means sensors connected to the Internet and behaving in an Internet-like way by making open, ad hoc connections, sharing data freely, and allowing unexpected applications, so computers can understand the world around them and become humanity's nervous system" (term delivered first in a 1999 oral presentation. See Ashton, 2015).

Our working definition is:

The IoT is a network of connected computing devices including different types of objects (e.g., digital machines). Each object in the network has a unique identifier (UID), and it is capable of collecting and transferring data automatically across the network.

The collected data has no value until it is analyzed, as illustrated in the opening vignette.

Note that the IoT allows people and things to interact and communicate at any time, any place, regarding any business topic or service.

According to Miller (2015), the IoT is a connected network in which:

- Large numbers of objects (things) can be connected.
- Each thing has a unique definition (IP address).
- Each thing has the ability to receive, send, and store data automatically.
- Each thing is delivered mostly over the wireless Internet.
- Each thing is built upon machine-to-machine (M2M) communication.

Note that, in contrast with the regular Internet that connects people to each other using computing technology, the IoT connects "things" (physical devices and people) to each other and to sensors that collect data. In Section 13.4, we explain the process of IoT.

SIMPLE EXAMPLES A common example of the IoT is the autonomous vehicle (Section 13.9). To drive on its own, a vehicle needs to have enough sensors that automatically monitor the situation around the car and take appropriate actions whenever necessary to adjust any setting, including the car's speed, direction, and so on. Another example that illustrates the IoT phenomenon is the company Smartbin. It has developed trash containers that include sensors to detect their fill levels. The trash collection company is automatically notified to empty a trash container when the sensor detects that the bin has reached the fill level.

A common example people give to illustrate IoT is the idea that a refrigerator could automatically order food (e.g., milk) when it detects that the food has run out! Clorox introduced a new Brita filter so that a Wi-Fi–enabled mechanism can order water filters by itself when it detects that it is time to change them. In these examples, a human does not have to communicate with another human or even with a machine.

IoT IS CHANGING EVERYTHING According to McCafferty (2015), the IoT is changing everything. This has been verified by a 2016 survey reported by Burt (2016). For how manufacturing is revolutionized by IoT, see Greengard (2016). Here are a few examples that he provided:

- "Real-time systems make it possible to know where anyone is at any moment, which is helpful to secured locations as military bases and seeking to push promotions to consumers."
- "Fleet tracking systems allow logistics and transport firms to optimize routing, track vehicle speeds and locations, and analyze driver and route efficiencies."
- "Owners and operators of jet engines, trains, factory equipment, bridges, tunnels, etc., can stay ahead of repairs through machines that monitor for preventive maintenance." (opening case)
- "Manufacturers of foods, pharmaceuticals and other products monitor temperature, humidity and other variables to manage quality control, receiving instant alerts when something goes wrong."

These changes are facilitated by AI systems, which enhance analytics and automate or support decision making.

The IoT Ecosystem

When billions of things are connected to the Internet with all the supporting services and connected IT infrastructure, we can see a giant complex, which can be viewed as a huge ecosystem. The **Internet of Things ecosystem** refers to all components that enable users to create IoT applications. These components include gateways, analytics, AI algorithms, servers, data storage, security, and connectivity devices. A pictorial view is provided in Figure 13.1 in which applications are shown on the left side and the building blocks and platforms on the right side. An example of an IoT application is provided in the opening vignette. It illustrates a network of sensors that collects information, which is transmitted to a central place for processing and eventually for decision support. Thus, the IoT applications are subsets of the IoT ecosystem.

A basic discussion, terms, major companies, and platforms is provided by Meola (2018).

Structure of IoT Systems

Things in IoT refers to a variety of objects and devices ranging from cars and home appliances to medical devices, computers, fitness tracers, hardware, software, data, sensors, and much more. Connecting things and allowing them to communicate is a necessary capability of an IoT application; but for more sophisticated applications, we need additional components: a control system and a business model. The IoT enables the things to sense or be sensed wirelessly across the network. A non-Internet example is a temperature control system in a room. Another non-Internet example is a traffic signal at intersections of roads where camera sensors recognize the cars coming from each direction and a control system adjusts the time for changing the lights according to programmed rules. Later, we will introduce the reader to many Internet-based applications.

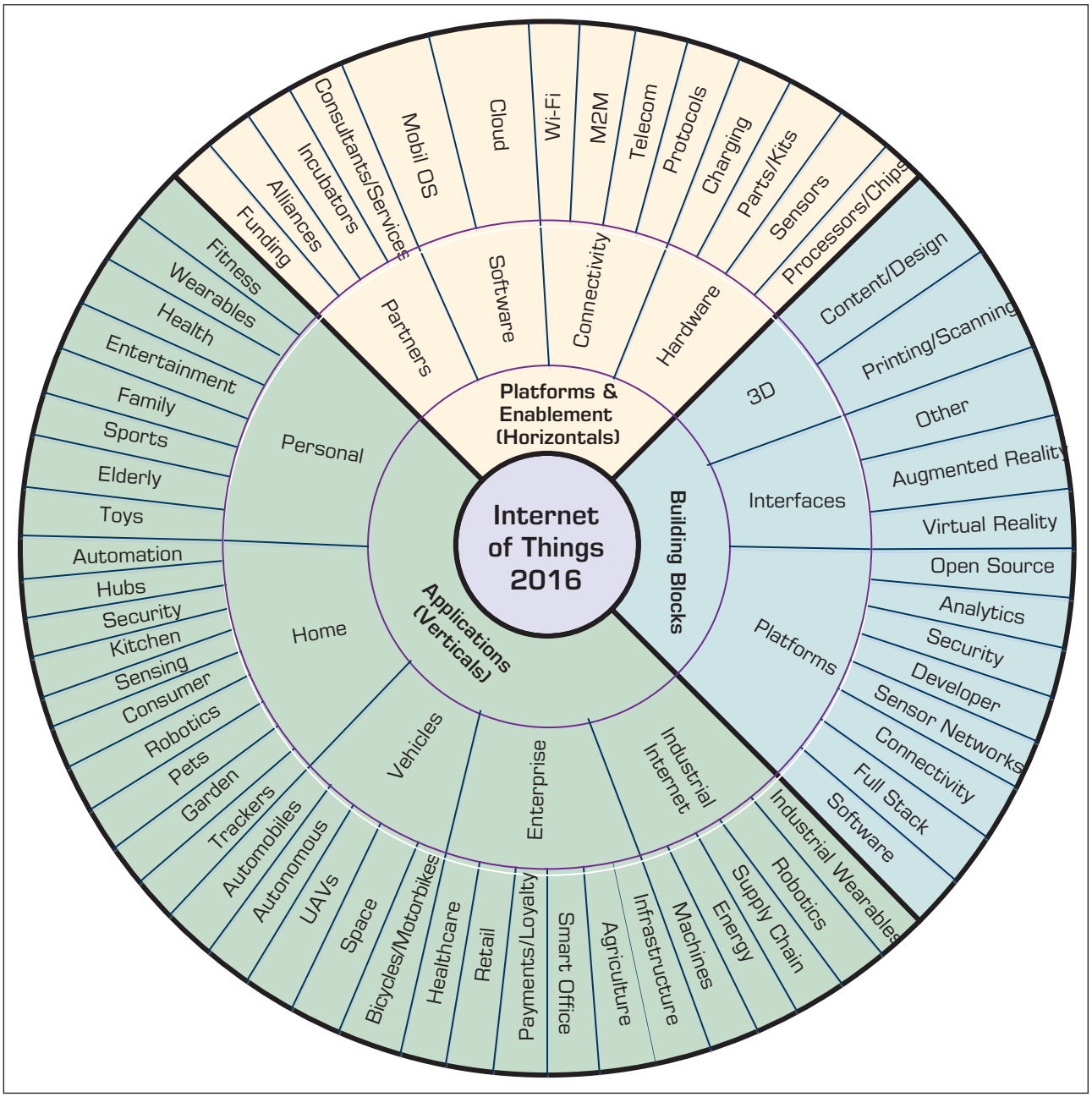


FIGURE 13.1 The IoT 2016 (Ecosystem).

IoT TECHNOLOGY INFRASTRUCTURE From a bird's-eye view, IoT technology can be divided into four major blocks. Figure 13.2 illustrates them.

1. ***Hardware:*** This includes the physical devices, sensors, and actuators where data are produced and recorded. The devices are the equipment that needs to be controlled, monitored, or tracked. IoT sensor devices could contain a processor or any computing device that parses incoming data.

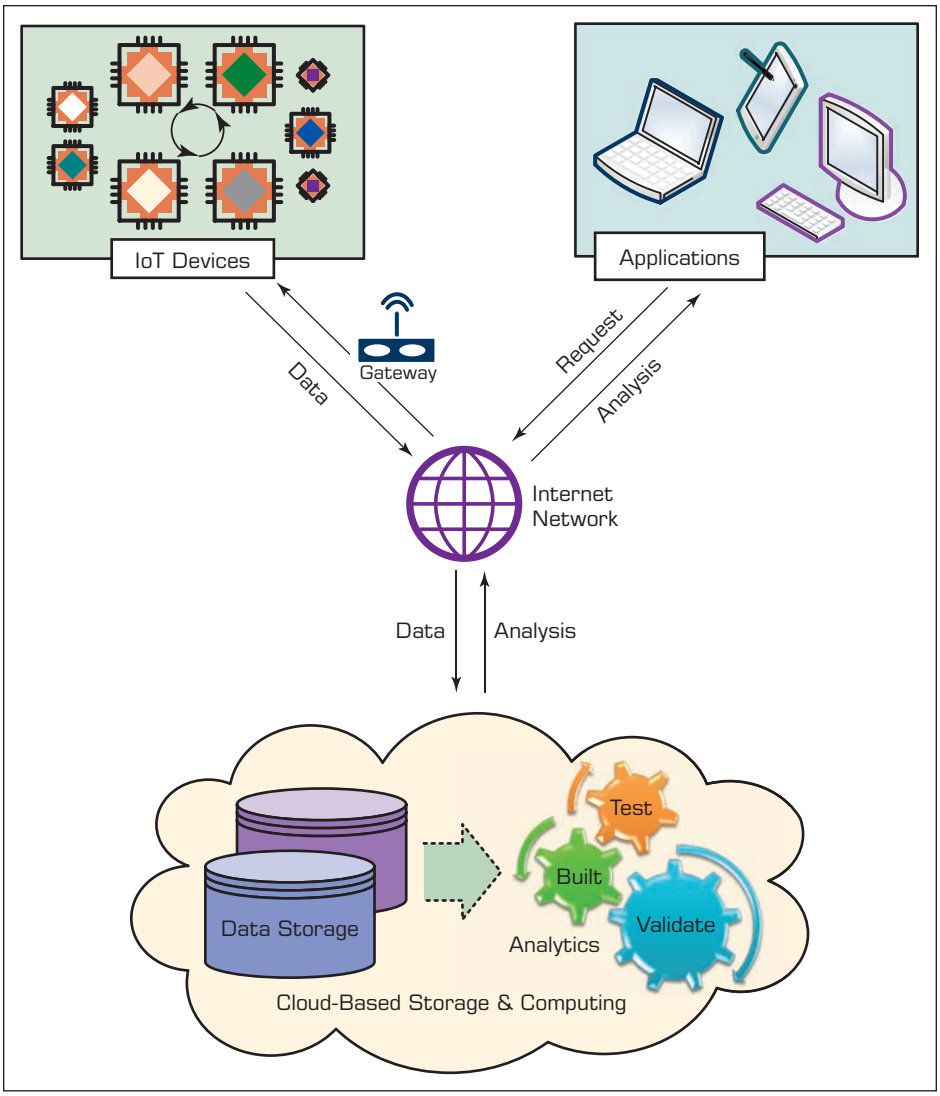


FIGURE 13.2 **The Building Blocks of IoT.**

2. ***Connectivity:*** There should be a base station or hub that collects data from the sensor-laden objects and sends those data to the "cloud" to be analyzed. Devices are connected to a network to communicate with other networks or other applications. These may be directly connected to the Internet. A gateway enables devices that are not directly connected to the Internet to reach the cloud platform.
3. ***Software backend:*** In this layer, the data collected are managed. Software back-end manages connected networks and devices and provides data integration. This may very well be in the cloud.
4. ***Applications:*** In this part of IoT, data are turned into meaningful information. Many of the applications can run on smartphones, tablets, and PCs and do something useful with the data. Other applications can run on the server and provide results or alerts through dashboards or messages to the stakeholders.

To assist with the construction of IoT systems, one may use IoT platforms. For information, see Meola (2018).

IoT PLATFORMS Because IoT is still evolving, many domain-specific and application-specific technology platforms are also evolving. Not surprisingly, many of the major vendors of IoT platforms are the same ones who provide analytics and data storage services for other application domains. These include Amazon AWS IoT, Microsoft Azure IoT suite, Predix IoT Platform by General Electric (GE), and IBM Watson IoT platform (**ibm.com/us-en/marketplace/internet-of-things-cloud**). Teradata Unified Data Architecture has similarly been applied by many customers in the IoT domain.

SECTION 13.2 REVIEW QUESTIONS

1. What is IoT?
2. List the major characteristics of IoT.
3. Why is IoT important?
4. List some changes introduced by IoT.
5. What is the IoT ecosystem?
6. What are the major components of an IoT technology?

13.3 MAJOR BENEFITS AND DRIVERS OF IoT

The major objective of IoT systems is to improve productivity, quality, speed, and the quality of life. There are potentially several major benefits from IoT, especially when combined with AI, as illustrated in the opening case. For a discussion and examples, see Jamthe, 2015.

Major Benefits of IoT

The following are the major benefits of IoT:

- Reduces cost by automating processes.
- Improves workers' productivity.
- Creates new revenue streams.
- Optimizes asset utilization (e.g., see the opening vignette).
- Improves sustainability.
- Changes and improves everything.
- May anticipate our needs (predictions).
- Enables insights into broad environments (sensors collect data).
- Enables smarter decisions/purchases.
- Provides increased accuracy of predictions.
- Identifies problems quickly (even before they occur).
- Provides instant information generation and dissemination.
- Offers quick and inexpensive tracking of activities.
- Makes business processes more efficient.
- Enables communication between consumers and financial institutions.
- Facilitates growth strategy.
- Fundamentally improves the use of analytics (see the opening vignette).
- Enables better decision making based on real-time information.
- Expedites problem resolution and malfunction recovery.
- Supports facility integration.
- Provides better knowledge about customers for personalized services and marketing.

Major Drivers of IoT

The following are the major drivers of IoT:

- The number of "things"—20 to 50 billion—may be connected to the Internet by 2020–2025.
- Connected autonomous "things"/systems (e.g., robots, cars) create new IoT applications.
- Broadband Internet is more widely available, increasing with time.
- The cost of devices and sensors is continuously declining.
- The cost of connecting the devices is decreasing.
- Additional devices are created (via innovations) and are interconnected easily (e.g., see Fenwick, 2016).
- More sensors are built into devices.
- Smartphones' penetration is skyrocketing.
- The availability of wearable devices is increasing.
- The speed of moving data is increasing to 60 THz.
- Protocols are developing for IoT (e.g., WiGig).
- Customer expectations are rising; innovative customer services are becoming a necessity.
- The availability of IoT tools and platforms is increasing.
- The availability of powerful analytics that are used with IoT is increasing.

Opportunities

The benefits and drivers just listed create many opportunities for organizations to excel in the economy (e.g., Sinclair, 2017), in many industries and in different settings.

McKinsey Global Institute (Manyika et al., 2015) provides a comprehensive list of settings where IoT is or can be used with examples in each setting. A 2017 study (Staff, 2017) revealed a dramatic increase in the capabilities and benefits of IoT.

HOW BIG CAN AN IoT NETWORK BE? While there will be billions of things connected to the Internet soon, not all of them will be connected in one IoT network. However, an IoT network can be very large, as we show next.

Example: World's Largest IoT Is Being Built in India (2017)

This network is being constructed by Tata Communications of India and HP Enterprises (HPE) of the United States, over the HPE Universal IoT Platform. The things to be connected exist in 2,000 communities and include computing devices, applications, and IoT solutions, connected over the Lo Ra network, a wireless communication protocol for wide area networks. The things are in smart buildings, utilities, university campuses, security systems, vehicles and fleets, and healthcare facilities.

The project is to be implemented in phases with proof-of-concept applications to be tested first. The network will bring services to 400 million people. For details, see Shah (2017).

SECTION 13.3 REVIEW QUESTIONS

1. List the benefits of IoT for enterprises.
2. List the benefits of IoT for consumers.
3. List the benefits of IoT for decision making.
4. List the major drivers of IoT.

13.4 HOW IoT WORKS

IoT is not an application. It is an infrastructure, platform, or framework that is used to support applications. The following is a comprehensive process for IoT applications. In many cases, IoT follows only portions of this process.

The process is explained in Figure 13.3. The Internet ecosystem (top of the figure) includes a large number of things. Sensors and other devices collect information from the ecosystem. The collected information can be displayed, stored, and processed analytically (e.g., by data mining). This analysis converts the *information* into *knowledge* and/or *intelligence*. Expert systems or machine learning may help in turning the *knowledge* into *decision support* (made by people and/or machines), which is evidenced by improved actions and results.

The generated decisions can help in creating innovative applications, new business models, and improvements in business processes These result in "actions," which may impact the original scenario or other things. The opening vignette illustrates this process.

Note that most of the existing applications are in the upper part of the figure, which is called *sensor to insight*, meaning up to the creation of knowledge or to the delivery of new information. However, now, the focus is moving to the entire cycle (i.e., *sensor to action*).

The IoT may generate a huge amount of data (Big Data) that needs to be analyzed by various business intelligence methods, including *deep learning,* or advanced AI methods.

IoT and Decision Support

As stated earlier, the IoT creates knowledge and/or intelligence, which is submitted as support to decision makers or is inputted to automated decision support entities. The transition from data collection to decision support may not be simple due to the large amount of data, some of which are irrelevant. Large-scale IoT usually needs to filter the

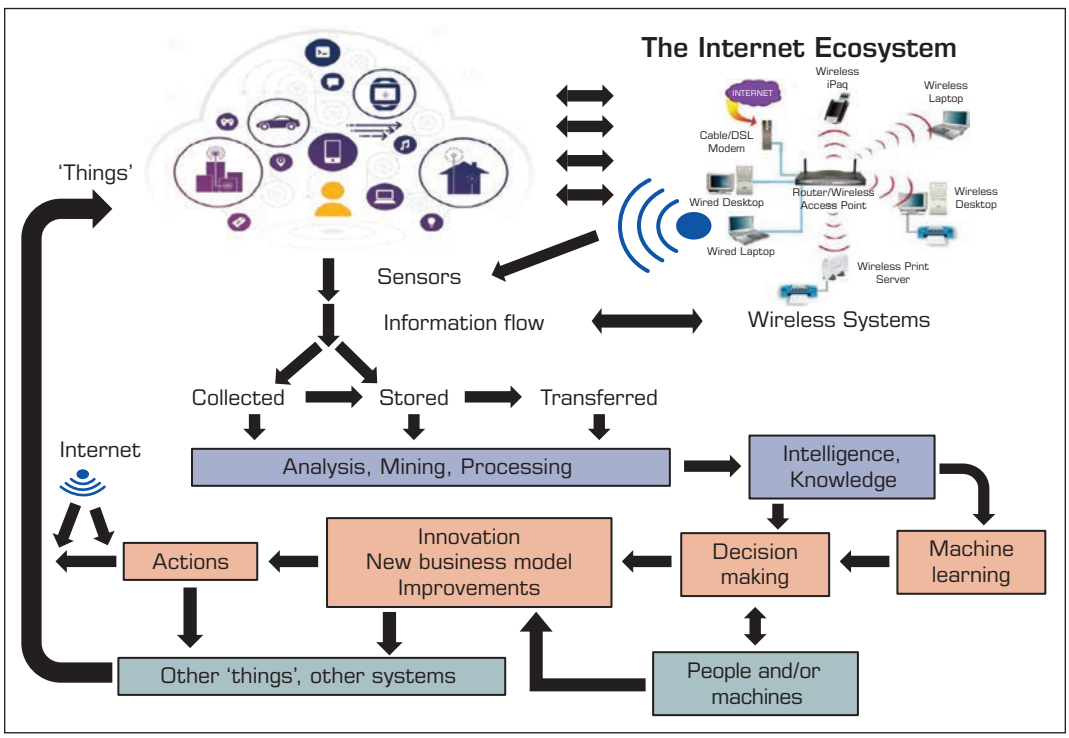


FIGURE 13.3 The Process of IoT.

collected data and "clean" them before they can be used for decision support, particularly if they are used as a base for automated decision making.

SECTION 13.4 REVIEW QUESTIONS

1. Describe the major components of IoT.
2. Explain how the IoT works following the process illustrated in Figure 13.3.
3. How does IoT support decision making?

13.5 SENSORS AND THEIR ROLE IN IoT

As illustrated in the opening vignette to this chapter, sensors play a major role in IoT by collecting data about the performance of the things that are connected to the Internet and monitoring the surrounding environment, collecting data there too if necessary. Sensors can transmit data and sometimes even process it prior to transmission.

Brief Introduction to Sensor Technology

A **sensor** is an electronic device that automatically collects data about events or changes in its environment. Many IoT applications include sensors (see the opening vignette). The collected data are sent to other electronic devices for processing. There are several types of sensors and several methods for collecting data. Sensors often generate signals that are converted to human-readable displays. In addition to their use in IoT, sensors are essential components in robotics and autonomous vehicles. Each sensor usually has a limit on the maximum distance that it can detect (nominal range). Sensors of a very short range known as *proximity sensors* are more reliable than those that operate in larger ranges. Each IoT network may have millions of sensors. Let us see how sensors work with IoT in Application Case 13.1.

Application Case 13.1 Using Sensors, IoT, and AI for Environmental Control at the Athens, Greece, International Airport

The Problem

Over 20 million passengers use the airport annually, and their number increases by more than 10 percent every year. Obviously, the number of flights is large and also increasing annually. The growth increases air pollution as well. The airport has a strong commitment to environmental protection, so management has looked for an environmental control solution. The objective was to make the airport carbon neutral. The large number of planes in the air and on the ground and the fact that airplanes frequently move require advanced technologies for the solution.

The Solution

A reasonable way to deal with moving airplanes was to use IoT, a technology that when combined with AI-based sensors enables environmental monitoring, analysis, and reporting, all of which provide the background information for decisions regarding minimizing the air pollution.

Two companies combined their expertise for this project: EXM of Greece, which specializes in IoT prediction analytics and innovative IoT solutions, and Libelium of the United States, which specializes in AI-related sensors, including those for environmental use. The objective of the project was to properly monitor air quality inside and outside the airport and to identify, in real time, the aircraft location on the ground and to take corrective actions whenever needed.

Ad Hoc Air Quality Monitoring and Analysis

The airport now has an air quality monitoring network. The solution includes Libelium's sensor

(*Continued*)

Application Case 13.1 (Continued)

platform connected in a cost-effective manner. The different sensors measure temperature, humidity, atmosphere pressure, ozone level, and particulate matter. The readings of the sensors are transmitted to IoT for reporting and then analysis. The sensors were improved by using AI features. Therefore, their accuracy increased. In addition, security and energy consumption are also being controlled.

Aircraft Location at the Airport

To identify the exact location of the aircrafts during takeoff and landing, the project uses acoustic measurement mechanisms. This is accomplished by using noise sensors placed in different locations. The sensors measure real-time noise level, which is evaluated by analytics. Overall, the system provides a noninvasive IoT solution.

Placement of sensors was difficult due to safety, security, and regulation considerations. Therefore, the sound monitoring subsystem had to be self-managed (autonomous), bearing solar panels and batteries that provided the electricity. In addition, the system utilizes a dual wireless communication system (known as GPPS).

The collected noise data are correlated with types of airplane and flights at the IoT backend. All data are analyzed by the airport environmental department and used for decisions regarding improvements of pollution control.

Technology Support

The solution combines an IoT system with AI-based analytics, visualization, and reporting and is executed in the cloud. In addition, the system has onsite sensors and communication infrastructures. Low-power wireless sensors monitor water and gas consumption indoors as well as air quality in the parking sites. Vendors' products, such as Microsoft Azure and IBM Bluemix, support the project and provide the necessary flexibility.

Sources: Compiled from Hedge (2017) and Twentyman (2017).

Questions for Case 13.1

1. What is the role of IoT in the project?
2. What is the role of sensors?
3. What are the benefits of the project?

How Sensors Work with IoT

In large-scale applications, sensors collect data that are transferred to processing in the "cloud." Several platforms are used for this process as discussed in Application Case 13.2.

Application Case 13.2 Rockwell Automation Monitors Expensive Oil and Gas Exploration Assets to Predict Failures

Rockwell Automation is one of the world's largest providers of industrial automation and information solutions. It has customers in more than 80 countries worldwide and around 22,500 employees. One of its business areas of focus is assisting oil and gas companies in exploration. An example is Hilcorp Energy, a customer company that drills oil in Alaska. The equipment used in drilling, extracting, and refining oil is very expensive. A single fault in the equipment can cost the company around $100,000 to $300,000 per day in lost production. To deal with this problem, it needed technology to monitor the status of such piece of equipment remotely and to predict failures that are likely to happen in the future.

Rockwell Automation considered the opportunity to expand its business in oil and gas industries by gathering data from the exploration sites and analyzing them to improve preventive maintenance decision making regarding the critical equipment, thus, minimizing downtime and drive better performance. The company utilizes its vision of Connected Enterprise with Microsoft's software

to monitor and support oil and gas equipment placed in remote areas. Rockwell is now providing solutions to predict failure of equipment along the entire petroleum supply chain, monitoring its health and performance in real time, and to prevent failures in the future. Solutions are provided in the following areas.

- **Drilling:** Hilcorp Energy has its pumping equipment stationed in Alaska where it drills for oil 24 hours a day. A single failure in equipment can cost Hilcorp a large amount of money. Rockwell connected electrical variable drives of pumping equipment to be processed in the "cloud," to control its machines thousands of miles away from the control room in Ohio. Sensors capture data, and through Rockwell's control gateway, these data are passed to Microsoft Azure Cloud. The solutions derived reach Hilcorp engineers through digital dashboards that provide real-time information about pressure, temperature, flow rate, and dozens of other parameters that help engineers monitor the equipment's health and performance. These dashboards also display alerts about any possible issues. When one of Hilcorp's pieces of pumping equipment failed, it was identified, tracked, and repaired in less than an hour, saving six hours of tracing the failure and the large cost of lost production.
- **Building smarter gas pumps:** Today, some delivery trucks use liquid natural gas (LNG) as fuel. Oil companies are updating their filling stations to incorporate LNG pumps. Rockwell Automation installed sensors and variable frequency drives at these pumps to collect real-time data about equipment operations, fuel inventory, and consumption rate. This data are transmitted to Rockwell's cloud platform for processing. Rockwell then generates interactive dashboards and reports using Microsoft Azure (an IoT platform). Results are forwarded to the appropriate stakeholders, giving them a good idea about the health of their capital assets.

The Connected Enterprise solution by Rockwell has accelerated growth for many oil and gas companies like Hilcorp Energy by bringing their operations data to the cloud platform and helping them reduce costly downtime and maintenance. It has resulted in a new business opportunity for industrial age stalwarts like Rockwell Automation.

Sources: **customers.microsoft.com** (2015); Rockwell Automation: *Fueling the Oil and Gas Industry with IoT*; **https://customers.microsoft.com/Pages/CustomerStory.aspx?recid=19922**; **Microsoft.com**. (n.d.). "Customer Stories | Rockwell Automation," **https://www.microsoft.com/en-us/cloud-platform/customer-stories-rockwell-automation** (accessed April 2018).

Questions for Case 13.2

1. What type of information would likely be collected by an oil and gas drilling platform?
2. Does this application fit the three V's (volume, variety, velocity) of Big Data? Why or why not?
3. Which other industries (list five) could use similar operational measurements and dashboards?

Sensor Applications and Radio-Frequency Identification (RFID) Sensors

There are many types of sensors. Some measure temperature; others measure humidity. Many sensors collect information and transmit it as well. For a list of 50 sensor applications with a large number of related articles, see **libelium.com/resources/top_50_iot_sensor_applications_ranking/**.

A well-known type of sensor that plays an important role in IoT is radio-frequency identification.

RFID SENSORS **Radio-frequency identification (RFID)** is part of a broader ecosystem of data capture technologies. Several forms of RFID in conjunction with other sensors play a major role in IoT applications. Let us see first what RFID is, as discussed in Technology Insights 13.1.

TECHNOLOGY INSIGHTS 13.1 RFID Sensors

RFID is a generic technology that refers to the use of radio-frequency waves to identify objects. Fundamentally, RFID is one example of a family of automatic identification technologies that also includes ubiquitous barcodes and magnetic strips. Since the mid-1970s, the retail supply chain (among many other areas) has used barcodes as the primary form of automatic identification. RFIDs can store a much larger amount of data than barcodes. Also, they can be accessed from a longer distance wirelessly. These potential advantages of RFID have prompted many companies (led by large retailers such as Walmart and Target) to aggressively pursue it as a way to improve their *supply chains* and thus reduce costs and increase sales. For details, see Chapter 8 in Sharda et al. (2018).

How does an RFID work? In its simplest form, an RFID system consists of a tag (attached to the product to be identified), an interrogator (i.e., RFID reader), one or several antennae attached to the reader, and a computer program (to control the reader and capture the data). At present, the retail supply chain has primarily been interested in using passive RFID tags. *Passive tags* receive energy from the electromagnetic field created by the interrogator (e.g., a reader) and backscatter information only when it is requested. The passive tag remains energized only while it is within the interrogator's magnetic field.

In contrast, *active tags* have a battery to energize themselves. Because active tags have their own power source, they do not need a reader to energize them; instead, they can initiate the data transmission process on their own. As compared to passive tags, active tags have a longer read range, better accuracy, more complex rewritable information storage, and richer processing capabilities. On the negative side, their batteries cause active tags to have a limited life span, be larger in size than passive tags, and be more expensive. Currently, most retail applications are designed and operated with passive tags, each of which costs only a few cents. Active tags are most frequently found in defense and military systems, yet they also appear in technologies such as EZ Pass whose tags (called *transponders*) are linked to a prepaid account that, for example, enables drivers to pay tolls later, by driving past a reader rather than stopping to pay at a tollbooth.

Note: There are also semipassive tags with limited active tag capabilities.

The most commonly used data representation for RFID technology is the Electronic Product Code (EPC), which is viewed by many in the industry as the next generation of the Universal Product Code (UPC), most often represented by a barcode. Like the UPC, the EPC consists of a series of numbers that identifies product types and manufacturers across the supply chain. The EPC also includes an extra set of digits to uniquely identify items.

Use of RFID and Smart Sensors in IoT

Basic RFID tags, either active or passive, are not sensors. The purpose of the tags is to identify objects and determine their location (e.g., for the purpose of counting objects). To make them useful for most IoT applications, the tags need to be upgraded (e.g., by adding on-board sensors). These RFIDs called *RFID sensors* have more capabilities than RFID tags, or basic sensors. For a detailed discussion about the role of RFID in the IoT, see Donaldson (2017).

RFID sensors are wireless sensors that communicate, via mash networks or conventional RFID readers, and they include identifiable ID. The RFID reader sends token information into gateways, such as AWS IoT service. This confirmation can be processed, resulting in some action.

SMART SENSORS AND IoT There are several types of smart sensors with different levels of capabilities when integrated into IoT. A **smart sensor** is one that senses the environment and processes the input it collects by using its built-in computing capabilities (e.g., a microprocessing). The processing is preprogrammed. Results are passed on. Depending on the internal computing quality, smart sensors can be more automated and accurate than other sensors and can filter out unwanted noise and compensate for errors before sending the data.

Smart sensors are crucial and an integral element in the IoT. They can include special components, such as amplifiers, analog filters, and transducers, to support IoT.

In addition, smart sensors for IoT can include special software for data conversion, digital processing, and communication capability to external devices.

According to a major study (Burkacky et al., 2018), sensors are getting smarter. Those on vehicles are examples. Vehicles can make the transition from being a hardware-driven machine to being a software-driven electronic device. Software can cost over 35 percent of the cost of vehicle production.

For further information, see Scannell (2017), Gemelli (2017), and Technavio (2017).

SECTION 13.5 REVIEW QUESTIONS

1. Define *sensor*.
2. Describe the role of sensors in IoT.
3. What is RFID? What is a RFID sensor?
4. What role does the RFID perform in IoT?
5. Define *smart sensor* and describe its role in IoT.

13.6 SELECTED IoT APPLICATIONS

We start with a well-known example: Imagine that your refrigerator can sense the amount of food in it and send you a text message when inventory is low (sensor-to-insight in Figure 13.3). One day refrigerators will also be able to place an order for items that need replenishment, pay for them, and arrange delivery (sensor-to-action). Let us look at some other, less futuristic enterprise applications.

A Large-scale IoT in Action

Existing contribution of IoT has centered on large organizations.

Example French National Railway System's Use of IoT

SNCF, the French national railway system, uses IoT to provide quality, availability, and safety for its nearly 14 million passengers. The company **sncf.com** improved its operations using IoT (Estopace, 2017a). To manage 15,000 trains and 30,000 kilometers of tracks is not simple, but IBM Watson, using IoT and analytics, helped to do just that. Thousands of sensors that are installed on the trains, tracks, and train stations gather data that Watson processes. In addition, all business process operations were digitized to fit into the system. Information concerning possible cyberattacks was also programmed into the system. All collected Big Data were prepared for decision support. IBM Watson's platform is scaleable and can handle future expansions.

To understand the magnitude of this IoT network, consider that the mass transit lines in Paris alone required 2,000 sensors forwarding information from more than 7,000 data points each month. The systems enable engineers to remotely monitor 200 trains at a time for any mechanical and electrical operations and malfunctions while trains are moving. In addition, by using a predictive analytic model, the company can schedule preventive maintenance to minimize failures. Therefore, if you are one of the train travelers, you can relax and enjoy your trip.

Examples of Other Existing Applications

The following examples of the use of IoT applications are based on information from Koufopoulos (2015):

- **Hilton Hotel.** Guests can check in directly to their rooms with their smartphones (no check-in lobby is needed, no keys are used). Other hotel chains follow suit.
- **Ford.** Users can connect to apps by voice. Autopaying for gas and preordering drinks at Starbucks directly from Ford's cars are in development.

- **Tesla.** Tesla's software autonomously schedules a valet to pick up a car and drive it to Tesla's facility when a car needs repair or schedule service. Tesla trucks, managed by IoT, will be driverless one day.
- **Johnnie Walker.** The whiskey company connected 100,000 of its bottles to the Internet for Brazil's Father's Day. Using smart labeling, buyers can create personalized videos to share with their fathers on social networks. Fathers also get promotions to buy more whiskey if they like it.
- **Apple.** Apple enables users of iPhones, Apple Watches, and Home kits to streamline shopping with Apple Pay.
- **Starbucks Clover Net in the Cloud.** This system connects coffee brewers to customers' preferences. It also monitors employee's performance, improves recipes, tracks consumption patterns, and so on.

A large number of consumer applications of IoT is reported by Jamthe (2016) and Miller (2015). For a list of IoT applications related to IBM Watson, see **ibm.com/internet-of-things/**.

Many companies are experimenting with IoT products for retailing (business to consumer, or B2C) and business to business (B2B) in areas such as operations, transportation, logistics, and factory warehousing. For the approaches of Apple and Amazon, see **appadvice.com/post/apple-amazons-smart-home-race/736365/**.

Note: For many case studies and examples of the IoT, see **ptc.com/en/product-lifecycle-report/services-and-customer-success-collide-in-the-iot**, **divante.co/blog/internet-e-commerce**, and Greengard (2016). IoT is also used for many applications inside enterprises (see McLellan, 2017a), and military purposes (see Bordo, 2016).

HOW IoT IS DRIVING MARKETING According to Durrios (2017), IoT can drive marketing opportunities in the following four ways:

1. *Disruptive data collection.* IoT collects more data about customers from more data sources than other technologies do. This includes data from wearables, smart homes, and everything consumers do. In addition, IoT provides data about changes in consumer preferences and behavior.
2. *Real-time personalization.* IoT can provide more accurate information about specific customers buying decisions, for example. IoT can identify customer expectations and direct customers to specific brands.
3. *Environmental attribution.* IoT can monitor environments regarding ad delivery for specific places, customers, methods, and campaigns. IoT can facilitate research of business environment; factors such competition, pricing, weather conditions, and new government regulations are observed.
4. *Complete conversation path.* IoT initiatives expand and enrich the digital channel of conversations between customers and vendors, especially those using wireless digital engagement. IoT also provides insight on consumer purchasing paths. In addition, marketers will receive improved customized market research data (e.g., by following the manner of customers' engagement and how customers react to promotions).

Of all the consumer-related IoT initiatives, three types are most well-known: smart homes and appliances (Section 13.7), smart cities (Section 13.8), and autonomous vehicles (Section 13.9). For more on IoT and customers, see Miller (2018).

SECTION 13.6 REVIEW QUESTIONS

1. Describe several enterprise applications.
2. Describe several marketing and sales applications.
3. Describe several customer service applications.

13.7 SMART HOMES AND APPLIANCES

The concept of the smart home has been in the limelight for several years, even before the concept of the IoT took a front stage.

A **smart home** is a home with automated components that are interconnected (frequently wirelessly), such as appliances, security, lights, and entertainment, and are centrally controlled and able to communicate with each other. For a description, see **techterms.com/definition/smart_home**.

Smart homes are designed to provide their dwellers with comfort, security, low energy cost, and convenience. They can communicate via smartphones or the Internet. The control can be in real time or at any desired intervals. Most existing homes are not yet smart, but they can easily and inexpensively be equipped to for at least partial smartness. Several protocols enable connections; well-known ones are XIO, UPB, Z-Wave, and EnOcean. These products offer scalability, so more devices can be connected to the smart home over time.

For an overview, see **techterms.com/definition/smart_home**, **smarthomeenergy.co.uk/what-smart-home**, and Pitsker (2017).

In the United States followed by other countries, thousands of homes are already equipped with such systems.

Typical Components of Smart Homes

The following are typical components in smart homes:

- **Lighting.** Users can manage their home lighting from wherever they are.
- **TV.** This is the most popular component.
- **Energy management.** Home heating and cooling systems can be fully automated and controlled via a smart thermostat (e.g., see **Nestnest.com/works-with-nest** about its product Nest Learning Thermostat).
- **Water control.** WaterCop (**watercop.com**) is a system that reduces water damage by monitoring water leaks via a sensor. The system sends a signal to a valve, causing it to close.
- **Smart speaker and chatbots** (see Chapter 12). Most popular are Echo and Alexa, and Google Assistant.
- **Home entertainment.** Audio and video equipment can be programmed to respond to a remote control device. For instance, a Wi-Fi–based remote control for a stereo system located in a family room can command the system to play on speakers installed anywhere else in the house. All home automation devices perform from one remote site and one button.
- **Alarm clock.** This tells kids to go back to sleep or to wake up.
- **Vacuum cleaner.** Examples are iRobot Roomba, and LG Roboking vacuum; see Chapter 2).
- **Camera.** This allows residents to see what is going on in their homes anytime from anywhere. Nest Cam Indoor is a popular product. Some smart cameras can even know how residents feel. See **tomsguide.com/us/hubble-hugo-smart-home-camera,news-24240.html**.

- **Refrigerator.** An example of this is Instaview from LG, which is powered by Alexa.
- **Home security and safety.** Such systems can be programmed to alert owners to security-related events on their property. As noted, some security can be supported by cameras for remote viewing of property in real time. Sensors can be used at home to detect intruders, keep an eye on working appliances, and perform several additional activities.

The major components of smart homes are illustrated in Figure 13.4.

Note that only a few homes have all of these components. Most common are home security, entertainment, and energy management.

Example: iHealthHome

Security measures are common in assisted living facilities in senior communities and for seniors who live independently. For example, the iHealthHome Touch screen system collects data and communicates with caregivers using the company's software. The system provides caregivers and physicians remote access to a person's health data. Using this technology, the iHealthHome program also reminds seniors of daily appointments and when to take their medicine. The system also reminds people when to self-measure their blood pressure and how to stay in touch with their caregivers.

Smart Appliances

A **smart appliance** includes features that can remotely control the appliance operations, based on the user preferences. A *smart appliance* may utilize a *Home Network* or the Internet to communicate with other devices in the smart home.

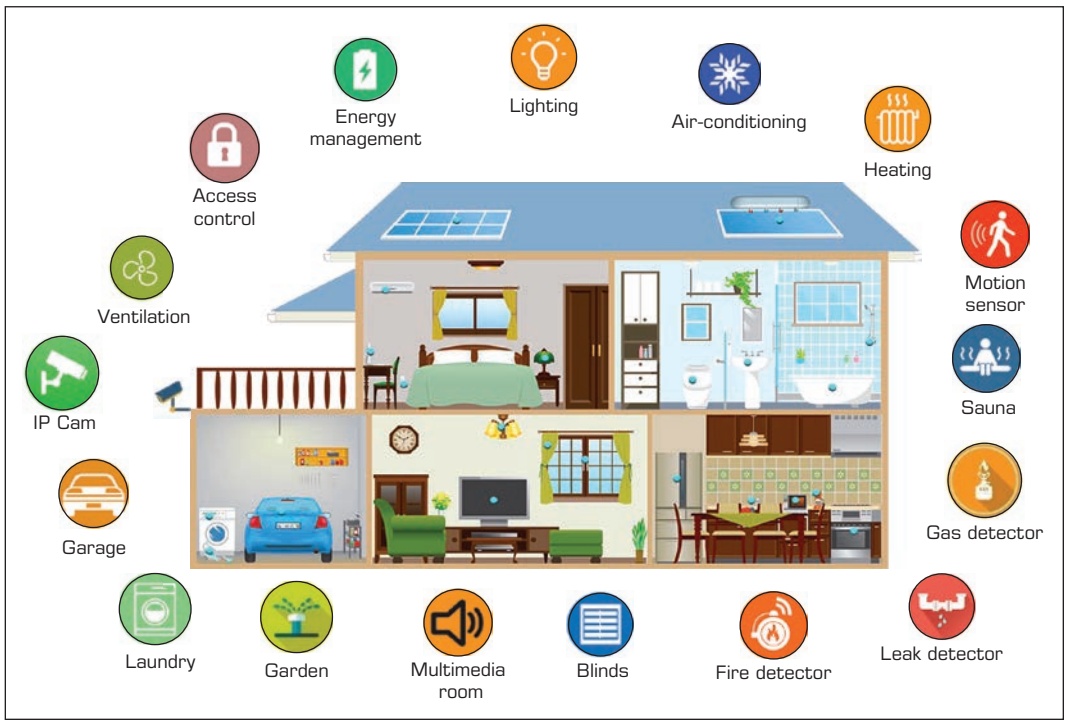


FIGURE 13.4 **The Components of a Smart Home.**

McGrath (2016) provides an overview of smart appliances that includes all appliances from Haier (a large China-based manufacturer). Its goal is to make everything in a house communicate across other device makers. Examples are smart refrigerators, air conditioners, and washing machines. Haier offers a control board for all appliances regardless of their manufacturers. Apple is working on a single control for all smart appliances in a home.

GOOGLE'S NEST A leading manufacturer of IoT smart home applications is Google's Nest. The company is a producer of programmable self-learning, sensor-driven, Wi-Fi–enabled products. In the spring of 2018, the company had three major products:

- **Learning thermostat.** This device learns what temperature and humidity level that people like and controls the air conditioner/heating system accordingly. Google claims that its products provide an average energy savings of 13 percent, which could pay for the device in two years; see **nest.com/thermostats/nest-learning-thermostat/overview/?alt=3**.
- **Smoke detector and alarm.** This device, which is controlled from a smartphone, tests itself automatically and lasts for about a decade. For details, see **nest.com/smoke-co-alarm/overview/**.
- **Nest.com.** This Webcam-based system allows users to see what is going on in their homes from any location via smartphone or any desktop computer. The system turns itself on automatically when nobody is at home. It can monitor pets, babies, and so on. A photo recorder allows users to go back in time. For details, see **nest.com/cameras/nest-cam-indoor/overview/**. For how Nest can use a phone to find out when individuals leave home, see Kastrenakes (2016). For more on Nest, see **en.wikipedia.org/wiki/Nest_Labs**.

Examples of Available Kits for Smart Homes

Two popular smart-home starter kits are (Pitsker, 2017):

1. ***Amazon Echo.*** This includes Amazon Echo, Belkin Wemo Mini, Philips Hue white starter kit, Ecobee Lite, and Amazon Fire TV stick with Alexa voice remote. Total cost on October 2017 was $495.
2. ***Google Home.*** This includes Google Home, Smart Speaker, Belkin Wemo Mini, Philips Hue white starter kit, Nest learning thermostat, and Google Chromecast (for entertainment). Total cost on October 2017 was $520.

HOME APPLIANCES IN CONSUMER ELECTRONIC SHOW (CES) 2016–2018 The following smart appliances, some of which were exhibited at the CES show in Las Vegas in January 2016 (Morris 2016), 2017, and 2018, are:

- **Samsung Smart fridge.** Cameras check content; sensors check temperature and humidity.
- **Gourmet robotic cooker.** It does interesting cooking.
- **10 in 1 device for the kitchen.** This stirs food such as scrambled eggs and has 10 cooking styles (e.g., baking, sauce making).
- **LG HUM-BOT Turbo+.** This can focus on an area in the home that needs special attention. A camera monitors the home remotely while the owner is away (similar to Google's Nest).

- **Haier R3D2 Refrigerator.** According to Morris (2016), this refrigeration is not the most practical one, but it has much of entertainment value. It looks like R3D2 in Star Wars. It can serve you a drink as well as provide lights and sounds.
- **Instaview Refrigerator from LG.** Powered by Alexa (enabled by voice), this includes a 29-inch LCD touch screen display. It provides functions such as determining the expiration dates of food and notifying the user. For details, see Diaz (2017).
- **Whirlpool's smart top load washer.** This fully automated machine has smart controls. It saves energy and even encourages philanthropy by sending a small amount of money to "Habitat for Humanity" each time washer is loaded.
- **LG LDT8786ST dishwasher.** This machine has camera whose sensors keep track of what has already been cleaned in order to save water. In addition, it provides flexibility in operations.

The following are smart home trends:

- TVs that can be used as a smart Hub for home appliances is coming from Samsung.
- Dolby Atmos products include speakers, receivers, and other entertainment items.
- DIY home smart security cameras make sure there is an intruder, not just the cat, before alerting the police.
- Water controls for faucets, sprinklers, and flood detectors are available. In addition, a robot can teach users how to save water indoors (hydrao.com/us/en/).

For more about home automation, see **smarthome.com/sh-learning-center-what-can-i-control.html**. Various apps used for home control can be found at **smarthome.com/android_apps.html**.

Smart components for the home are available at home improvement stores (e.g., Lowes) and can be purchased directly from manufacturers (e.g., Nest).

To facilitate the creation of smart components for the home, Amazon and Intel Corp. partnered in 2017 to provide developers with platforms to advance the smart home ecosystem. For details, see **pcmag.com/news/350055/amazon-intel-partner-to-advance-smart-home-tech/**.

For smart appliances at CES 2018, watch the video at **youtube.com/watch?v=NX-9LivJh0/**.

A Smart Home Is Where the Bot Is

The virtual personal assistant that we introduced in Chapter 12 enables people to converse by voice with chatbots such as Alexa/Echo and Google Assistant. Such assistants can be used to manage appliances in smart homes.

In a comprehensive smart home, devices not only meet household needs but also are able to anticipate them. It is predicted that in the near future, an AI-based smart home will feature an intelligent and coordinated ecosystem of bots that will manage and perform household tasks and may even be emotionally connected with people. For a prediction of the future bots, see Coumau et al. (2017). Amazon and Intel joined forces to develop such smart home ecosystems that include NLP capabilities.

Smart homes will also have smart robots that can serve people snacks, help take care of people who are handicapped, and even teach children different skills.

Barriers to Smart Home Adoption

The potential of smart homes is attractive, but it will take some time before there will be many of them. The following are some limiting barriers, per Vankatakrishnan (2017).

- ***Compatibility.*** There are too many products and vendors to choose from, making potential buyers confused. Many of these products do not "speak" to each other, so more industry standards are needed. In addition, it is difficult to match the products with consumers' needs.
- ***Communication.*** Different consumers have different ideas on what the smart home should be. Therefore, the capabilities and benefits of a smart home need to be clearly communicated to users.
- ***Concentration.*** Brands need to concentrate on population segments that are most interested in smart homes (e.g., Gen Y).

In addition are the issues of cost justification, invasion of privacy, security, and ease of use. For the future of smart homes, including the role of Amazon and Walmart, and how the smart home will shop for itself, see Weinreich (2018).

Smart homes, appliances, and buildings can be featured in smart cities, the subject of our next section.

SECTION 13.7 REVIEW QUESTIONS

1. Describe a smart home.
2. What are the benefits of a smart home?
3. List the major smart appliances.
4. Describe how Nest works.
5. Describe the role of bots in smart homes.

13.8 SMART CITIES AND FACTORIES

The idea of smart cities took off around 2007 when IBM launched its Smart Planet project and Cisco began its Smart Cities and Communities program. The idea is that in **smart cities**, digital technologies (mostly mobile based) facilitate better public services for citizens, better utilization of resources, and less negative environmental impact. For resources, see **ec.europa.eu/digital-agenda/en/about-smart-cities**. Townsend (2013) provides a broad historical look and coverage of the technologies. In an overview of his book, he provides the following examples: "In Zaragoza, Spain, a 'citizen card' can get you on the free city-wide Wi-Fi network, unlock a bike share, check a book out of the library, and pay for your bus ride home. In New York, a guerrilla group of citizen-scientists installed sensors in local sewers to alert you when storm water runoff overwhelms the system, dumping waste into local waterways." According to a prediction made by Editors (2015), smart cities would use 1.6 billion connected things in 2016. Finally, smart cities can have several smart entities such as universities and factories (see Lacey, 2016). For more on smart cities, see Schwartz (2015). In addition, watch the video "Cisco Bets Big on 'Smart Cities'" at **money.cnn.com/video/technology/2016/03/21/cisco-ceo-smart-cities.cnnmoney**. Another video to watch is "Smart Cities of the Future" (3:56 min.) at **youtube.com/watch?v=mQR8hxMP6SY**. A more detailed video on San Diego (44:06 minutes) is at **youtube.com/watch?v=LAjznAJe5uQ**.

Cities cannot become smart overnight, as illustrated in Application Case 13.3, which presents the case of Amsterdam and its evolution into a smart city.

In many countries, governments and others (e.g., Google) are developing smart city applications. For example, India has begun to develop 100 smart cities (see **enterpriseinnovation.net/article/india-eyes-development-100-smart-cities-1301232910**).

Application Case 13.3 Amsterdam on the Road to Become a Smart City

In over seven years, the city of Amsterdam (The Netherlands) was transformed into a smart city using information technologies. This case describes the steps the city took from 2009 to 2016 to become a smart city, as reported by MIT Sloan School of Management. The city initiative included projects in the following categories: mobility, quality of living, transportation, security, health, and economy as well as infrastructure, big and open source data, and experimental living labs.

The major findings of the MIT team regarding Amsterdam's transformation were:

- **Private-sector data are critical for changing policy.** The major categories of the project involved nongovernmental entities (e.g., using a GPS provider to manage traffic). For example, the private sector was involved in a project to change traffic situations (reduction of 25 percent in the number of cars and an increase of 100 percent in the number of scooters, in five years).
- **It is necessary to have chief technology officers in smart cities.** Smart cities require the collection of large amounts of data using several tools and algorithms. Issues such as cost and security are critical.
- **Expectations of the contribution of the IoT, Big Data, and AI, need to be managed.** Citizens expect rapid changes and improvement in areas ranging from parking to traffic. Data collection is slow, and changes are difficult to implement.
- **Smart city initiatives must start with data inventory.** The problem in Amsterdam was that data were stored in 12,000 databases across 32 departments. These were organized differently on different hardware, so data inventory was needed. This initial activity was boring and tedious and had no immediate visible payoff.
- **Pilot projects are an excellent strategy. Pilot projects provide lessons for future projects.** The city had over 80 pilot projects, for example, collecting different types of trash and placing them in different colored bags. Successful projects are scaled up in size.
- **Citizen input is a critical success factor.** There are several ways to encourage citizens to provide input. Involvement of universities and research institutions is also critical. In addition, social media networks can be used to facilitate citizens' engagement.

The smart city initiative may be only in its beginning, but it is already improving the quality of life of residents and increasing the economic growth of the city. A critical success factor of the initiative was the willingness of the city officials to share their data with technology companies.

IoT was a major component in the projects. First, it enabled the flow of data from sensors and databases for analytic processing. Second, IoT enables autonomous vehicles of all kinds, which contribute to the reduction of pollution, vehicle accidents, and traffic jams. Finally, IoT provides real-time data that help decision makers develop and improve policies. In April 2016, the city won Europe's "Capital of Innovation" award (a prize of 950,000 euros).

Sources: Compiled from Brokaw (2016), Fitzgerald (2016), **amsterdamsmartcity.com**, and **facebook.com/amsterdamsmartcity.**

Questions for Case 13.3

1. Watch the video at **youtube.com/watch?v=FinLi65Xtik/** and comment on the technologies used.
2. Get a copy of the MIT case study at **sloanreview.mit.edu/case-study/data-driven-city-management/**. List the steps in the process and the applications that were likely used in IoT.
3. Identify the smart components used in this project.

Smart Buildings: From Automated to Cognitive Buildings

IBM'S COGNITIVE BUILDINGS In a white paper (IBM, 2016), IBM discussed the use of IoT to make *cognitive buildings*, which are able to learn the behavior of a building's system in order to optimize it. The cognitive building does so by autonomously integrating the IoT devices with the IoT operation. Such integration enables the creation of new business processes and increases the productivity of existing systems. Based on the concept of cognitive computing (Chapter 6), IBM describes the maturity of the technology as a continuation of the phase that started with *automated buildings* (1980 to 2000), the creation of *smart building* (2000 to 2015), and finally, cognitive building (beginning in 2015). The process is illustrated in Figure 13.5. The figure also shows the increased capabilities of buildings over time.

The highlights of a cognitive building are:

- By applying advance analytics, buildings can provide insights in near real time.
- It learns and reasons from data and interacts with humans. The system can detect and diagnose abnormal situations and propose remedies.
- It has the ability to change building temperature subject to humans' preferences.
- It is aware of its status and that of its users.
- It is aware of its energy status and adjusts it to be comfortable to dwellers.
- Its users can interact with the building via text messages and voice chatting.
- Robots and drones are starting to operate inside and outside the building without human intervention.

A major collaborator of IBM is Siemens (from Germany). The companies concentrate on global issues related to the use of IoT to enhance building performance.

Smart Components in Smart Cities and Smart Factories

The major objective of smart cities is to automate as many as possible public services such as transportation, utilities, social services, security, medical care, education, and economy. So, in the smart city overall project one may find several subprojects, some of which are independent of the master project.

Example

Hong Kong has a project called a *smart mobility* for the improvement of road safety. A consortium of private and public organizations has introduced Intelligent Transport

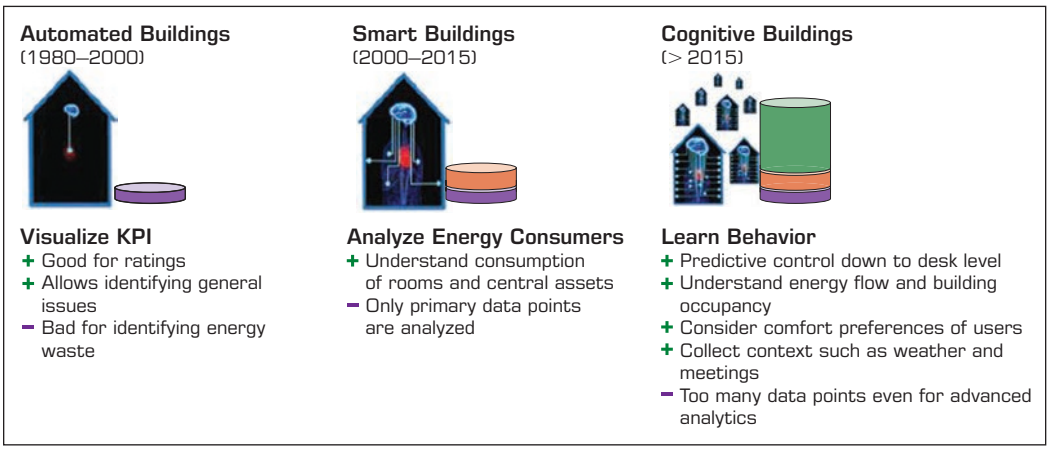


FIGURE 13.5 **IBM's Cognitive Building Maturity Framework.** *Source:* IBM. "Embracing the Internet of Things in the new era of cognitive buildings." IBM Global Business Services, White Paper, 2016. Courtesy of International Business Machines Corporation, © International Business Machines Corporation.Used with permission.

Services, including a warning mechanism for collision, and control assistance for finding parking. The system also manages speed and lane violations and traffic congestion. All of these increase safety and efficiency. For details, see Estopace (2017b).

Transportation is a major area in which analytics and AI can make cities smarter. Other areas include economic development, crime fighting, and healthcare. For details, see SAS (2017).

Other examples of smart city components can be found in a smart university, smart medical centers, smart power grid, and in airports, factories, ports, sport arenas, and smart factories. Each of these components can be treated as an independent IoT project, and/or as a part of the smart city overall project.

SMART (DIGITAL) FACTORIES Automation of manufacturing has been with us for generations. Robots are making thousands of products from cars to cellphones. Tens of thousands of robots can be found in Amazon's distribution centers. Therefore, it is not surprising that factories are getting smarter with AI technologies and IoT applications. As such they may be considered a component of smart cities and may be interrelated with other components, such as clean air and transportation.

A **smart factory**, according to Deloitte University Press, is "a flexible system that can self-optimize performance across a broader network, self-adapt to and learn from new conditions in real or near real time, and autonomously run entire production processes." For details, see the free Deloitte e-book at **DUP_The-smart-factory.pdf**. For a primer, see **https://www2.deloitte.com/insights/us/en/focus/internet-of-things/technical-primer.html**.

Tomás (2016) provides a vision of what industrial production will look like in the future. It will be essentially fully digitized and connected, fast, and flexible. The major idea is that there will be a command center in a factory equipped with AI technologies. The AI, combined with IoT sensors and information flow, will enable optimal organization and sequencing of business processes. The entire production chain, from raw material suppliers, logistics, and manufacturing to sales, will be connected to IoT systems for planning, coordination, and control. Planning will be based on analytic predictions of demand.

Production processes will be automated as much as possible and wirelessly controlled. Logistics will be provided on demand quickly, and quality control will be automated. IoT combined with sensors will be used for both predictive and preventive maintenance. Some of these elements exist in advanced factories, and more factories will be smarter in the future.

For more on smart factories, see Libelium (2015) and Pujari (2017). For the smart factory of the future, read **belden.com/blog/industrial-ethernet/topic/smart-factory-of-the-future/page/0**.

The use of IoT in the factory is illustrated in the video "Smart Factory Towards a Factory of Things" at **youtube.com/watch?v=EUnnKAFcpuE** (9:10 min.).

Smart factories will have different business processes, new technology solutions, different people-machine interactions, and a modified culture. For the transformation process to a smart factory, see Bhapkar and Dias (2017). The accounting firm Deloitte (**dupress.deloitte.com/smart-factory**) provides a diagram that illustrates "the major characteristics of a smart factory" (Figure 13.6).

Example: Smart Bike Production in a Smart Factory

The world demand for smart bikes is increasing rapidly, especially in smart cities. Mobike is the world's first and largest bike-sharing company. To meet the demand, the company is working with Foxconn Technology Group to make the bike production smarter. The smart manufacturing involves the creation of a global supply chain from raw materials to production to sales. Foxconn is known for its high-technology expertise in

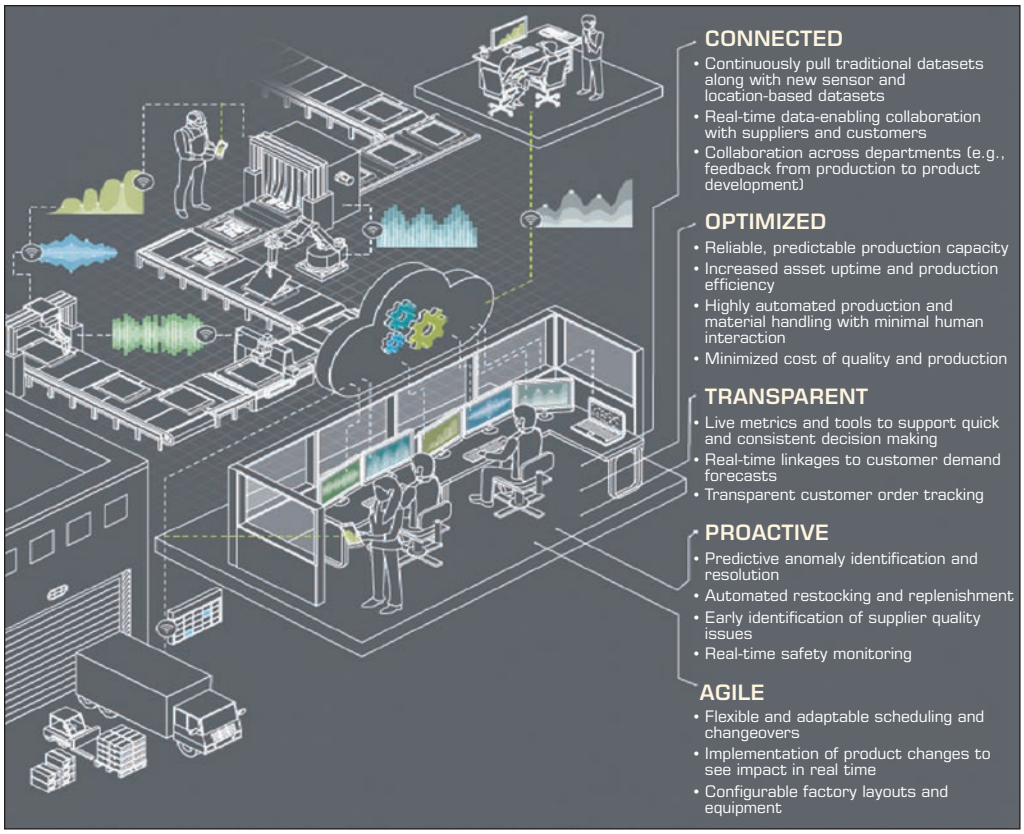


FIGURE 13.6 **Five Key Characteristics of a Smart Factory (Deloitte).** *Source:* Burke, Hartigan, Laaper, Martin, Mussomeli, Sniderman, "The smart factory: Responsive, adaptive, connected manufacturing," Deloitte Insights (2017), **https://www.deloitte.com/insights/us/en/focus/industry-4-0/smart-factory-connected-manufacturing.html**. Used with permission.

providing efficient manufacturing processes in a cost-efficiency production. It optimizes Internet-driven smart manufacturing. The production output is expected to double in the near future. For details, see Hamblen (2016) and **enterpriseinnovation.net/article/foxconn-drives-mobike-smart-bike-production-1513651539**.

EXAMPLES OF SMART CITY INITIATIVES Smart city initiatives are diversified, as explained earlier. For examples, see Application Case 13.4.

Application Case 13.4 How IBM Is Making Cities Smarter Worldwide

IBM has been supporting smart city initiatives for several years. The following examples are compiled from Taft's slide show (**eweek.com/cloud/how-ibm-is-making-cities-smarter-worldwide**).

- ***Minneapolis (United States).*** The initiative supports more effective decisions for the city's resource allocation. In addition, it aligns the operations of multiple departments working on the same project. IBM is providing AI-based pattern recognition algorithms for problem solving and performance improvement.
- ***Montpellier (France).*** IBM's software is helping the city in its initiatives of water manage-

(*Continued*)

Application Case 13.4 (Continued)

ment, mobility (transportation), and risk management (decision making). The rapidly growing city must meet the increasing demand for services. To do this efficiently, IBM provides data analysis and interpretation of activities, research institutions, and other partners in the region.

- ***Stockholm (Sweden).*** To reduce traffic problems, IBM technologies are optimally matching demands and supplies. The initiative uses sensors and IoT to alleviate the congestion problem.
- ***Dubuque (United States).*** Several initiatives were conducted for efficient use of resources (e.g., utilities) and management of transportation problems.
- ***Cambridge (Canada).*** The city is using IBM's "Intelligent Infrastructure Planning" for conducting business analytics and decision support technologies. Using AI-based algorithms, the city can make better decisions (e.g., repair or replace assets). In addition, IBM smart technologies help to improve project coordination.
- ***Lyon (France).*** Transportation management is a major project in any big city and a target for most smart city initiatives. Smart technologies provide transportation staff with effective real-time decision support tools. This helped reduce traffic congestion. Using predictive analytics, future problems can be forecasted, so, if they occur, they can be solved quickly.
- ***Rio de Janeiro (Brazil).*** To manage and coordinate the operations of 30 city departments is a complex undertaking. IBM technologies support a central command center for the city that plans operations and handles emergencies in all areas.
- ***Madrid (Spain).*** To manage all its emergency situations (fire, police transportation, hospitals), the city created a central response center. Data are collected by sensors, GPS, surveillance cameras, and so on. The center was created after Madrid's 2004 terrorist attack and is managed with the support of IBM smart technologies.
- ***Rochester (United States).*** The city police department is using IoT and predictive analysis to forecast when and where crimes is likely to be committed. This AI-based system has proven to be accurate in several other cities.

These examples illustrate the utilization of IBM's Smarter Cities framework in several areas by smart city initiatives. Note that IBM Watson is using IoT for many of its own projects.

Questions for Case 13.4

1. List the various services that are improved by IoT in a smart city.
2. How do the technologies support decision making?
3. Comment on the global nature of the examples.

A major area of improvement in a smart city is transportation.

Improving Transportation in the Smart City

A major problem in many cities is the increased number of vehicles and the inability to accommodate all of them effectively. Building more roads could add more pollution and lead to traffic jams. Public transportation can help alleviate the problem but may take years to complete. Quick solutions are needed. In the opening case to Chapter 2, we introduced Inrix. The Inrix company uses AI and other tools to solve transportation problems. It collects data from stationary sensors along roads and from other sources. In some smart cities, innovators have already placed air quality sensors on bicycles and cars. Sensors also are taking data from cars on the roads to help generate data that can analyzed and results are transmitted to drivers. An example of another innovative project is provided in the following examples.

Example 1

Valerann, an Israeli start-up, developed smart road studs to replace the reflective studs of today's technology. Smart studs can transmit information of what they sense about what

is occurring on the roads. Eventually, the studs will be incorporated with autonomous vehicles. The smart studs cost more than reflective studs but have a longer life. For details, see Solomon (2017).

Example 2

Smart Mobility Consortium (Hong Kong) works on mobility in the smart city of Hong Kong. More than 10 million people there use the public and private transportation systems every day. This transportation project includes several smart subsystems for parking, collision warning, and alerts for speeders and lane changing violators. For details, see Estopace (2017b).

Combining Analytics and IoT in Smart City Initiatives

Like in many IoT initiatives, it is necessary to combine analytics and IoT. A notable example is IBM Watson. Another one is the SAS platform.

Example: The SAS Analytics Model for Smart Cities

The amount of data collected by IoT networks in cities can be enormous. Data are collected from many sensors, computer files, people, databases, and so on. To make sense of these data, it is necessary to use analytics, including AI algorithms. SAS is using a seven-step process divided into three major phases: *Sense, Understand*, and *Act*. The following are definitions of these (condensed from SAS, 2017).

- *Sense*. Using sensors, sense anything that matters. SAS analyzes the collected data. The data go through intelligent filters for cleanliness so that only relevant data go to the next phase. IoT collects and transfers the data from the sensors.
- *Understand the signals in the data*. Using data mining algorithms, the entire relevant ecosystem is analyzed for pattern recognition. The process can be complex as the data collected by IoT sensors are combined with data from other sources.
- *Act*. Decisions can be made quickly as all relevant data are in place. SAS decision management tools can support the process. Decisions range from alerts to automated actions.

The SAS process is illustrated in Figure 13.7. For more on analytics and IoT combination, see SAS Analytics for IoT at **https://www.sas.com/en_us/insights/big-data/internet-of-things.html**. For additional information, see Henderson (2017).

Bill Gates' Futuristic Smart City

In November 2017, Bill Gates purchased 60,000 acres of land west of Phoenix, Arizona, where he plans to construct a futuristic city from scratch. The city will be a model and place for research.

Technology Support for Smart Cities

A large number of vendors, research institutions, and governments are providing technology support for smart cities. Here are few examples.

TECHNOLOGY SUPPORT BY BOSCH CORP. AND OTHERS Bosch Corp (of Germany), a major supplier of automotive parts, presented several innovations related to smart cities at CES 2018.

According to Editors (2018), revenues of global smart cities with IoT technology will exceed $60 billion by 2026.

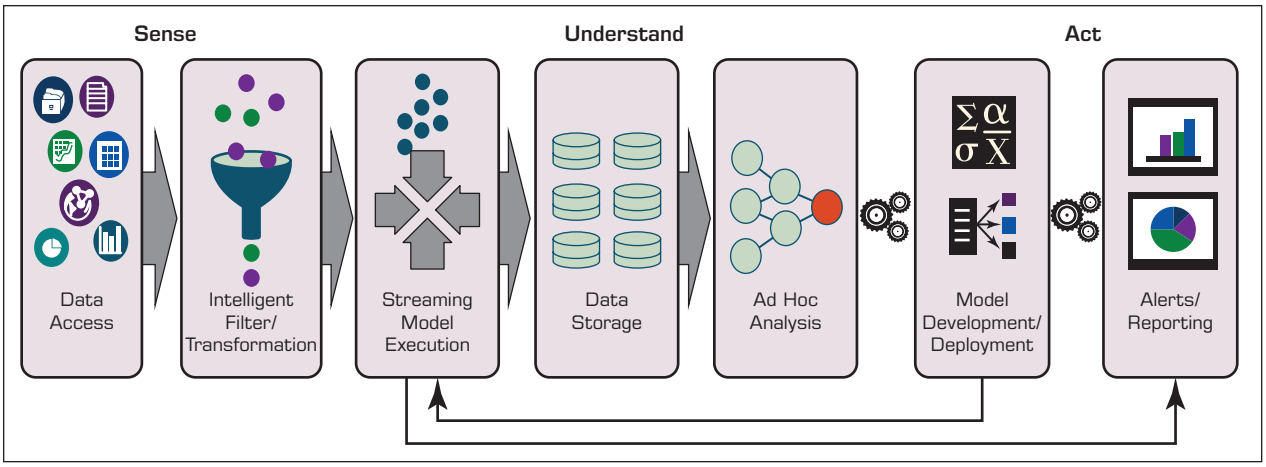


FIGURE 13.7 SAS Supports the Full IoT Analytics Life Cycle for Smart Cities (SAS). *Source:* Courtesy of SAS Institute Inc. Used with permisison.

Finally, in smart cities, connected and self-driven vehicles will be everywhere (see Hamblen, 2016 and the next section).

SECTION 13.8 REVIEW QUESTIONS

1. Describe smart city.
2. List some benefits of a smart city to the residents.
3. What is the role of IoT in smart city initiatives?
4. How are analytics combined with IoT? Why?
5. Describe smart and cognitive buildings.
6. What is a smart factory?
7. Describe technology support to smart cities.

13.9 AUTONOMOUS (SELF-DRIVING) VEHICLES

Autonomous vehicles, also known as **driverless cars**, robot-driven cars, self-driving cars, and autonomous cars, are already on the roads in several places. The first commercial autonomous car project was initiated by Google (named Google Chauffeur) and is becoming a reality, with several U.S. states preparing to allow them on the road. France, Singapore, China, and several other countries already have these cars and buses on their roads. These cars are electric, and they can create a revolution by reducing emissions, accidents, fatalities (an estimate of about 30,000 fatalities a year, worldwide), and traffic jams (e.g., see Tokuoka, 2016). Thus far, these cars are being tested in several cities worldwide and in some cities are already on the roads. Experts estimate that 10 million such cars will be on the roads in the United States by 2020, and China is planning for 30 million cars by 2021.

The Developments of Smart Vehicles

The initial efforts to commercialize a self-driving car were started by Google in the 1990s. These efforts can be seen today in Waymo's story in Application Case 13.5.

Application Case 13.5 Waymo and Autonomous Vehicles

Waymo is a unit of Alphabet (previously called Google) that is fully dedicated to the Google self-driving car project. Almost 20 years ago, Google, with the help of Stanford University, started to work on this project. The idea received a boost in 2005 when DARPA awarded its Grand Challenge prize to the project. Then, the U.S. Department of Defense awarded it a $2 million prize. Google pioneered physical experiments in 2009 after conducting computer simulation for several years when it ran self-driving cars 2.5 billion virtual miles. The next step was to get legislation to allow autonomous vehicles on the roads. By 2018, 10 states had passed such laws. Some allow robot-driven cars only in certain areas. Self-driving cars (see a Waymo car in Figure 13.8) with robot-only chauffeurs were tested in early 2018 by Waymo in the Phoenix, Arizona, area. First, corporate engineering will be in the driver's seat; but, around November 2018, the cars were expected to be completely driverless. The company was ready to start running commercial minivans in five states in 2018. By the end of 2018, Waymo vans were expected to pick up regular passengers who volunteered to take the service (called Early Rider Program), although most travelers are still skeptic.

This works in the following way. Company technicians, acting like regular riders, order service via a mobile app. The AI mechanism figures out how the vehicle will get to the requested caller as well as how it will self-drive to the requested destination.

Waymo, the pioneer of autonomous vehicles, collaborated with Chrysler (using Chrysler Pacifica minivans). The computing power is provided by Intel (with its Mobileye division). The high cost of the cars will limit their use initially to commercial uses. However, Waymo already has agreed to manage Avis's fleet of self-driving minivans. Also, realizing the power of ride-sharing services, Waymo is working with Lyft on new autonomous vehicles. Finally, Waymo is partnering with AutoNation to provide maintenance and road services for Waymo cars.

Note: On the legal dispute involving Uber, see the opening case of Chapter 14.

Sources: Compiled from Hawkins (2017), Ohnsman (2017), and Khoury (2018).

Questions for Case 13.5

1. Why did Waymo first use simulation?
2. Why was legislation needed?
3. What is the Early Rider Program?
4. Why will it take years before regular car owners will be able to enjoy a ride in the back seat of their self-driving cars?
5. Why are Lyft, Uber, and Avis interested in self-driving cars?

FIGURE 13.8 Waymo (Google) Self-Driving Car. *Source:* SiliconValleyStock/Alamy Stock Photo.

An example of how Nvidia works with Toyota's initiative is presented in Technology Insights 13.2.

TECHNOLOGY INSIGHTS 13.2 Toyota and Nvidia Corp. Plan to Bring Autonomous Driving to the Masses

It is not surprising that Toyota is interested in smart cars. As a matter of fact, the company's cars are expected to be on the market in 2020. Toyota plans to produce several types of autonomous vehicles. One type will be for elderly and disabled people. Another type will have the ability to drive completely autonomously or be an assistant (with a mechanism called "guardian angel") to drivers. For example, it will have the ability to take full control when the driver falls asleep, or when it senses that an accident is coming. A tired driver will be able to use Alexa (or a similar device) to tell the guardian angel to take over.

Autonomous vehicles need a smart control system, and this is where Nvidia enters the picture. Autonomous cars need to process a vast amount of data collected by sensors and cameras in real time. Nvidia pioneered a special AI-based supercomputer (called Drive PX2) for this purpose. The computer includes a special processor (called Xavier) that can power the autonomous driving gear of the cars. The partnership with Toyota enables Nvidia to leverage the power of its processor to apply AI to the autonomous cars.

Nvidia's supercomputer has an AI algorithm-based special operating system that includes a cloud-based 3D map with high definition. With these capabilities, the car's "brain" can comprehend its driving surroundings. Since a car can also exactly identify its own location, it will know about any potential hazard (e.g., road work or a vehicle coming toward it). The operating system is being constantly updated, so it makes the car smarter (AI learning capability).

The Xavier system provides the car's "brain" on a special chip (called Volta), which can deliver 30 trillion *deep learning* operations per second. Thus, it can process complex AI algorithms involving machine learning. Nvidia is expected to use Volta to open a new, powerful era in AI computing.

Source: Compiled from Korosec (2017) and **blogs.nvidia.com/blog/2016/09/28/Xavier/**.

Questions for Discussion

1. What does a car need to have in order to be autonomous?
2. What is the contribution of Nvidia to self-driving cars?
3. What is the role of Xavier?
4. Why does the process use a supercomputer?

Despite the required complex technology, several car manufacturers are ready to sell or operate such cars soon (e.g., BMW, Mercedes, Ford, GM, Tesla, and of course—Google).

Developments related to driverless vehicles follow:

- Uber and other ride-sharing companies plan for self-driving cars.
- Mail is delivered to homes by self-driving cars; see **uspsoig.gov/blog/no-driver-needed**.
- Driverless buses are being tested in France and Finland. Watch **money.cnn.com/video/technology/2016/08/18/self-driving-buses-hit-the-road-in-helsinki.cnnmoney** about self-driving buses in Helsinki.
- Self-driving taxis already operate in Singapore.

The Self Drive Act is the first national law in the United States pertaining to self-driving cars. It aims to regulate the safety of the passengers in autonomous vehicles. It opens the door for the production of 100,000 cars per year by 2021.

Flying Cars

While autonomous vehicles on the road may have considerable difficulties, there is research on flying cars. As a matter of fact, drones that can carry people already exist. As long as there is not much traffic in the air, there will be no traffic problem. However, the navigation of a large number of flying cars may be a problem. Airbus created a flying taxi demo in 2016 and Uber developed the concept and summarized it in a 98-page report released in October 2016. Toyota is also working on making a flying car. In January 2018, at the Las Vegas CES, Intel showed an autonomous passenger drone named Volocopter. This machine can be developed as an air taxi one day. For flying taxis in New Zealand, see Sorkin (2018).

Implementation Issues in Autonomous Vehicles

Autonomous vehicles such as cars, trucks, and buses are already on the roads in several cities worldwide. However, before we will see millions of them on the roads, it will be necessary to deal with several implementation issues. The following are reasons why full commercialization is going to take time:

- The cost of real-time 3D map technologies needs to be reduced and their quality needs to be increased.
- AI software must be nimble and its capabilities increased. For example, AI needs to deal with many unexpected conditions, including that of the behavior of drivers of other cars.
- Bray (2016) posted an interesting question: "Are customers, automakers and insurers really ready for self-driving cars?" Customers seem to acknowledge that such cars are coming. But they resist boarding one. However, some daring people expect these cars to do a better job than humans in driving.
- The technology needs more research, which is very expensive. One reason is that the many sensors in the cars and on the road need to be improved and their cost need to be reduced.
- The IoT is connecting many objects for autonomous vehicles, including those in clouds. The IoT systems themselves need to be improved. For example, data transmission delays must be eliminated. For more IT/AI generic implementation issues, see Chapter 14.

➤ SECTION 13.9 REVIEW QUESTIONS

1. What are self-driving vehicles? How are they related to the IoT?
2. What are the benefits of self-driving vehicles to drivers, society, and companies?
3. Why are Uber and similar companies interested in self-driving vehicles?
4. What AI technologies are needed to support autonomous vehicles?
5. What are flying cars?
6. List some implementation issues of autonomous vehicles.

13.10 IMPLEMENTING IoT AND MANAGERIAL CONSIDERATIONS

In this chapter, we presented a number of successful IoT-based applications. The results so far are more than encouraging, especially in areas such as monitoring equipment performance to improve its operation and maintenance (e.g., CNH in the opening vignette and the IBM Watson case of elevators in Chapter 1). However, this is only the tip of the iceberg. As we indicated earlier, the IoT can change everything. In this section,

we present some of the major issues that are related to successful IoT implementation. Although there is considerable excitement about the growth and the potential of the IoT, there are that managers should be aware of.

Major Implementation Issues

McKinsey's Global Institute (Bughin et al., 2015) has put together a comprehensive *Executive's Guide to the Internet of Things*. This guide identifies the following issues:

- **Organizational alignment.** Although it is true of several other technology initiatives, with IoT, the opportunities for operational improvements and creating new business opportunities means that IT and operational personnel have to work as one team rather than separate functions. As noted by the guide's authors, "IoT will challenge other notions of organizational responsibilities. Chief financial, marketing, and operating officers, as well as leaders of business units, will have to be receptive to linking up their systems."
- **Interoperability challenges.** Interoperability is a huge detriment thus far in the growth of IoT applications. Few IoT devices connect seamlessly with each another. Second, there are many technological issues regarding connectivity. Many remote areas do not yet have proper Wi-Fi connection. Issues related to Big Data processing are also responsible for slow progress in IoT adoption. Companies are trying to reduce data at the sensor level so that only a minimal amount goes into clouds. Current infrastructure hardly supports the huge amount of data collected by IoT. A related problem is retrofitting sensors on devices to be able to gather and transmit data for analysis. In addition, it will take time for consumers to replace their analog objects with new IoT digital smart products. As an example, it is easier for people to replace mobile phones than a car, kitchen appliances, and other things that can benefit from having a sensor and being connected to IoT.
- **Security.** Security of data is an issue in general, but it is an even bigger one in the context of IoT. Each device that is connected to IoT becomes another entry point for malicious hackers to get into a large system or at the least operate or corrupt a specific device. There are stories of hackers being able to breach and control automated functions of a car or to control a garage door opener remotely. Such issues require that any large-scale adoption of IoT involve security considerations from the very beginning.

Given that the Internet is not well secured, applying IoT networks requires special security measures, especially in the wireless sections of the networks. Perkins (2016) summarizes the situation as follows: "IoT creates a pervasive digital presence connecting organizations and society as a whole. New actors include data scientists; external integrators; and exposed endpoints. Security decision makers must embrace fundamental principles of risk and resilience to drive change." For a free e-book about IoT, see McLellan (2017b).

Additional issues follow.

- **Privacy.** To ensure privacy, one needs a good security system plus a privacy protection system and policy (see Chapter 14). Both may be difficult to construct in IoT networks due to the large size of the networks and the use of the less protected Internet. For advice from top security experts, see Hu (2016).
- **Connection of the silos of data.** There are millions of silos of data on the Internet and many of them need to be interconnected in specific IoT applications. This issue is known as the need for a "fabric" and connectivity. This can be a complex issue for applications that involve many different silos belonging to different organizations. Connectivity is needed in machine to machine, people to

people, people to machines, and people to services and sensors. For a discussion, see Rainie and Anderson (2017) and **machineshop.io/blog/the-fabric-of-the-internet-of-things**. For how the connection is done at IBM Watson, see **ibm.com/Internet-of-things/iot-solutions/**.

- **Preparation of existing IT architectures and operating models for IoT can be a complex issue in many organizations.** For a complete analysis and guide on this subject, see Deichmann et al. (2015). Integrating IoT into IT is critical for the data flow needed by the IoT and IoT-processed data to flow back to actions.
- **Management.** As in the introduction of any new technology, the support of top management is necessary. Bui (2016) recommends hiring a *chief data officer* in order to succeed in IoT due to the need to deal with silos of data described earlier. Using such a top manager can facilitate information sharing across all business functions, roles, and levels. Finally, it solves departmental struggles to own and control the IoT.
- **Connected customers.** There is evidence of an increased use of IoT in marketing and customer relationships. In addition, the IoT drives increased customer engagement. According to Park (2017), a successful deployment of IoT for customers requires "connected customers." The connection needs to be for data, decisions, outcomes, and staff related to any contacts relevant to the IoT and marketing. The Blue Hill research organization provides a free report on this issue (see Park). IoT enables a better connection with key clients and improves customer service. Of special considerations are hospitality, healthcare, and transportation organizations.

Finnaly, Chui et al. (2018) provided suggestions in a recent study on how to succeed in IoT implementation.

With so many implementation issues, an implementation strategy is necessary.

Strategy for Turning Industrial IoT into Competitive Advantage

IoT collects large amounts of data that can be used to improve external business activities (e.g., marketing) as well as internal operations. SAS (2017) proposed a strategy cycle that includes the following steps:

1. *Specify the business goals.* They should be set with perceived benefits and costs so the initiatives can be justified. This step involves a high level of planning and examination of resources. Initial return on investment (ROI) analysis is advisable.
2. *Express an analytic strategy.* To support ROI and prepare a business case, it will be necessary to plan how Big Data will be analyzed. This involves the selection of an analytic platform, which is a critical success factor. An examination of emerging AI technologies, such as deep learning, may be conducted. An appropriate selection will ensure a powerful IoT solution.
3. *Evaluate the needs for edge analytics.* Edge analytics is a technology that is needed for some, but not all, applications. It is designed to introduce real-time capabilities to the applications. It also filters data to enable automated decision making, frequently in real time because only relevant data results from the filtering.
4. *Select appropriate analytics solutions.* There are numerous analytic solutions on the market offered by many vendors. In using one or several for IoT, it is necessary to consider several criteria such as fitness for IoT, ease of deployment, ability to minimize project risks, sophistication of the tools, and connection to existing IT systems (e.g., the quality of IoT gateways). Sometimes it is a good idea to look at a group of vendors that offer combined products (e.g., SAS and Intel). Finally, appropriate infrastructures, such as high-performance cloud servers and storage systems, need to be examined. These must work together as a scalable, effective, and efficient platform.

5. *Continues improvement closes the loop.* Like in any strategy cycle, performance should be monitored, and improvements in various steps of the process need to be considered, especially since IoT is evolving and changing rapidly. The extent of goal achievement is an important criteria and upgrading the goals should be considered.

A summary of the process is provided in Figure 13.9.

Weldon (2015) suggests the following steps for successful IoT implementation:

- Develop a business case to justify the IoT project including a cost-benefit analysis and a comparison with other projects.
- Develop a working prototype. Experiment with it. Learn and improve it.
- Install the IoT in one organizational unit; experiment with it. Learn lessons.
- Plan an organization-wide deployment if the pilot is a success. Give special attention to data processing and dissemination.

The Future of the IoT

With the passage of time, we see an increasing number of IoT applications, both external and internal to organizations and enterprises. Because all IoT networks are connected to the Internet, it will be possible to have some of the networks connected to each other, creating larger IoTs. This will create growth and expansion opportunities for many organizations.

AI ENHANCEMENT OF IoT There are several areas of potential development. One area where AI will enhance IoT is in its ecosystem. Many IoT applications are complex and could be improved with machine learning that can provide insights about data. In addition, AI can help in creating devices ("things") that can self-diagnose problems and even repair them. For further discussion, see Martin (2017). Another future benefit of AI when combined with IoT is "shaping up to be a symbiotic pairing" (Hupfer, 2016). This pairing can create cognitive systems that are able to deal with and understand data that

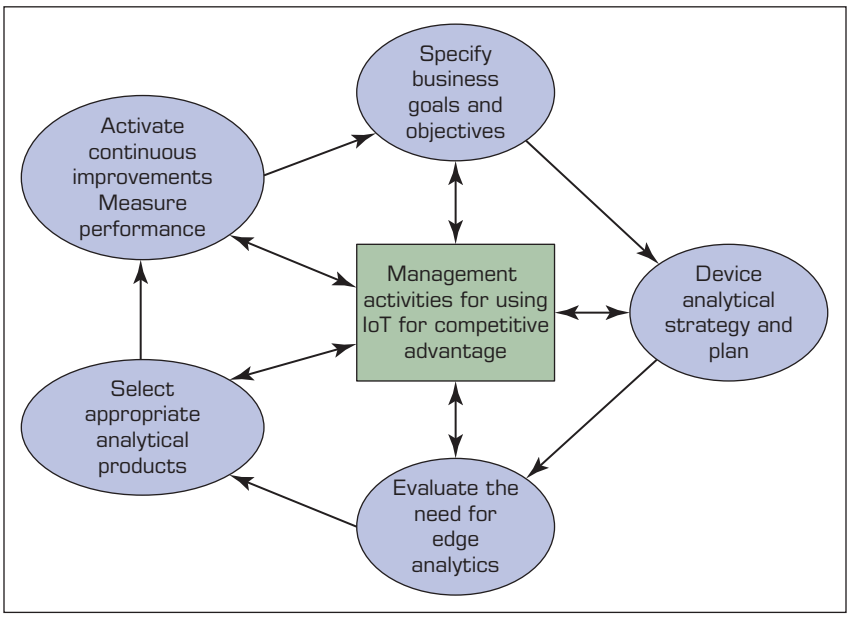


FIGURE 13.9 The IoT Strategy Cycle.

conventional analytics cannot handle. The AI and IoT combination can create an *embodied cognition* that injects AI capabilities into objects (such as robots and manufacturing machines) to enable the objects to understand their environments and then self-learn and improve their operation. For details, see Hupfer (2017). Finally, AI can help the integration of IoT with other IT systems.

A final word! By now you are probably interested to know about getting a job in IoT. Yes, there is a shortage of IoT experts, and annual salaries can range from $250,000 to $500,000. For 2017 data, see Violino (2017).

Chapter Highlights

- The IoT is a revolutionary technology that can change everything.
- The IoT refers to an ecosystem in which a large number of objects (such as people, sensors, and computers) are interconnected via the Internet (frequently wirelessly). By the years 2020 to 2025, there could be as many as 50 billion connected objects. Subsystems of such connected things can be used for many purposes.
- Use of the IoT can improve existing business processes and create new business applications.
- Billions of things will be connected to the Internet, forming the IoT ecosystem.
- Things on the IoT will be able to communicate, and the structure will enable a central control to manipulate things and support decision making in IoT applications.
- The IoT enables many applications in industry, services, and governments.
- IoT applications are based on analysis of data collected by sensors or other devices that flow over the Internet for processing.
- Sensors can collect data from a large number of things (e.g., over 1 million elevators in the opening case of Chapter 1).
- Major efforts are needed to connect the IoT with other IT systems.
- IoT applications can support decisions made by equipment manufacturers and by the users of equipment. (See the opening vignette of this chapter.)
- IBM Watson is a major provider of IoT applications in many industries and services (e.g., medical research). It was projected to reach over 1 billion users by the end of 2018.
- Smart appliances and homes are enabled by IoT.
- Smart city projects worldwide are supported by IoT, increasing the quality of life for residents of the cities and supporting the decision making of city planners and technology providers.
- Self-driven cars may reduce accidents, pollution, traffic jams, and transportation costs. Self-driving cars are not fully implemented yet, but some were introduced in 2018.
- Smart homes and appliances are popular. For a small cost, owners can use several applications from home security to controlling appliances in their homes.
- The concept of smart cities is being developed globally with projects in countries such as India, Germany, and the United States and the city-state of Singapore. The objective of smart cities is to provide a better life for their residents. Major areas covered are transportation, healthcare, energy saving, education, and government services.

Key Terms

autonomous vehicles (driverless cars)
Internet of Things
Internet of Things ecosystem
radio-frequency identification (RFID)
sensor
smart appliance
smart cities
smart factory
smart homes
smart sensors

Questions for Discussion

1. Compare the IoT with regular Internet.
2. Discuss the potential impact of autonomous vehicles on our lives.
3. Why must a truly smart home have a bot?
4. Why is the IoT considered a disruptive technology?
5. Research Apple Home Pod. How does it interact with smart home devices?
6. Alexa is now connected to smart home devices such as thermostats and microwaves. Find examples of other appliances that are connected to Alexa and write a report.
7. Discuss the objective of smart cities to conserve the earth's limited resources.
8. What are the major uses of IoT?
9. Accidents involving driverless cars slow down the implementation of the technology. Yet, the technology can save hundreds of thousands of lives. Is the slowdown (usually driven by politicians) justifiable? Discuss.

Exercises

1. Go to theinternetofthings.eu and find information about the *IoT Council*. Write a summary of it.
2. Go to **https://www.ptc.com/en/resource-center** or other sources, and select three IoT implemented cases. Write a summary of each.
3. AT&T is active in smart city projects. Investigate their activities (solutions). Write a summary.
4. It is said that the IoT will enable new customer service and B2B interactions. Explain how.
5. The IoT has a growing impact on business and e-commerce. Find evidence. Also read Jamthe (2016).
6. Find information about Sophia, a robot from Hanson Robotics. Summarize her capabilities.
7. Examine the Ecobee thermostat and its integration with Alexa. What are the benefits of the integration? Write a report.
8. Enter **smartcitiescouncil.com**. Write a summary of the major concept found there; list the major enablers and the type of available resources.
9. Find the status of Bill Gates's futuristic smart city. What are some of its specific plans?
10. *City Brain* is the name of Alibaba's platform for smart cities. One project has been adopted in China and Malaysia. Find information and write a report.
11. Find the status of delivering pizza by self-driving cars. Check Domino's Pizza news.
12. India has many IoT applications, including projects for 100 smart cities. Read the 2016 status report **atenterpriseinnovation.net/article/internet-things-next-big-wave-india-1270947471/** and find more recent information about it. Why do you think IoT is so widespread in India? Write a report.
13. Read the Blue Hill report (Park, 2017) and summarize all the issues related to IoT.
14. Find the status of *smart cities* as it is related to IoT and Cisco. Write a report.
15. Watch the video **atyoutube.com/watch?v=ZJr0X3XBMmA** (14:36 min.). Write a summary about the five smart devices.
16. Watch the video "Smart Manufacturing" (22 min.) at **youtube.com/watch?v=SfVUkGoCA7s** and summarize the lessons learned.
17. The competition for creating and using autonomous cars is intensifying globally. Find 12 companies that are competing in this field.
18. Enter McKinsey Global Institute **mckinsey.com/mgi/overview** and find recent studies on IoT. Prepare the summary.
19. AT&T is trying to connect autonomous vehicles to smart cities. Find information on the progress of this project. Identify the benefits and the difficulties.

References

Ashton, K. *How to Fly a Horse: The Secret History of Creation, Invention and Discovery*. New York City, NY: Doubleday, January 2015.

Bhapkar, R., and J. Dias "How a Digital Factory Can Transform Company Culture." *McKinsey & Company*, September 2017.

Bordo, M. "Israeli Air Force Works on Battlefield IoT Technology." *ReadWrite.com*, June 21, 2016.

Bray, E. "Are Consumers, Automakers and Insurers Really for Self-Driving Cars?" *Tech Crunch*, August 10, 2016.

Brokaw, L. "Six Lessons from Amsterdam's Smart City Initiative." *MIT Sloan Management Review*, May 25, 2016.

Bughin, J., M. Chui, and J. Manyika. "An Executive's Guide to the Internet of Things." *McKinsey Quarterly*, August 2015.

Bui, T. "To Succeed in IoT, Hire a Chief Data Officer." *Tech Crunch*, July 11, 2016.

Burkacky, O., et al. "Rethinking Car Software and Electronics Architecture." *McKinsey & Company*, February 2018.

Burt, J. "IoT to Have Growing Impact on Businesses, Industries, Survey Finds." *eWeek*, May 4, 2016.

Chui, M., et al. "What It Takes to Get an Edge in the Internet of Things?" *McKinsey Quarterly*, September 2018.

Coumau, J., et al. "A Smart Home Is Where the Bot Is." *McKinsey Quarterly*, January 2017.

Deichmann, J., M. Roggendorf, and D. Wee. "Preparing IT Systems and Organizations for the Internet of Things." *McKinsey & Company*, November 2015.

Diaz, J. "CES 2017: LG's New Smart Fridge Is Powered by Alexa." *Android Headlines*, January 4, 2017. **androidheadlines.com/2017/01/ces-2017-lgs-new-smart-fridge-powered-alexa.html/** (accessed August 2018).

Donaldson, J. "Is the Role of RFID in the Internet of Things Being Underestimated?" *Mojix*, May 2, 2017.

Durrios, J. "Four Ways IoT Is Driving Marketing Attribution." *Enterprise Innovation*, April 8, 2017.

Editors. "Smart Cities Will Use 1.6B Connected Things in 2016." *eGov Innovation*, December 22, 2015.

Editors. "Global Smart Cities IoT Technology Revenues to Exceed US$60 Billion by 2026." *Enterprise Innovation*, January 23, 2018.

Estopace, E. "French National Railway Operator Taps IoT for Rail Safety." *eGov Innovation*, February 21, 2017a.

Estopace, E. "Consortium to Build a Smart Mobility System for Hong Kong." *Enterprise Innovation*, March 26, 2017b.

Fenwick, N. "IoT Devices Are Exploding on the Market." *Information Management*, January 19, 2016.

Fitzgerald, M. "Data-Driven City Management: A Close Look at Amsterdam's Smart City Initiative." *MIT Sloan Management Review*, May 19, 2016.

Freeman, M. "Connected Cars: The Long Road to Autonomous Vehicles." *San Diego Union Tribune*, April 3, 2017.

Gemelli, M. "Smart Sensors Fulfilling the Promise of the IoT." *Sensors Magazine*, October 13, 2017.

Greengard, S. "How AI Will Impact the Global Economy." *CIO Insight*, October 7, 2016.

Hamblen, M. "Smart City Tech Connects Cars and Bikes with Big Data at MCW: Innovators Can Put Air Quality Sensors on Bicycles, While Wireless Connections Help Pave the Way for Driverless Cars." *Computerworld*, February 22, 2016.

Hawkins, A. "Intel Is Working with Waymo to Build Fully Self-Driving Cars." *The Verge*, September 18, 2017.

Hedge, Z. "Case Study: Athens International Airport Uses EXM and Libelium's IoT Platform to Enhance Environmental Monitoring." *IoT Now*, September 1, 2017.

Henderson, P. "10 Ways Analytics Can Make Your City Smarter." *InfoWorld* and *SAS Report AST* = 0182248, June 6, 2017.

Hu, F. *Security and Privacy in Internet of Things (IoTs): Models, Algorithms, and Implementations*. Boca Raton, FL: CRC Press, 2016.

Hupfer, S. "AI Is the Future of IoT." *IBM Blog*, December 15, 2016. **ibm.com/blogs/internet-of-things/ai-future-iot/** (accessed July 2018).

IBM. "Embracing the Internet of Things in the New Era of Cognitive Buildings." White Paper. *IBM Global Business Services*, 2016.

Jamthe, S. *The Internet of Things Business Primer*. Stanford, CA: Sudha Jamthe, 2015.

Jamthe, S. *IoT Disruptions 2020: Getting to the Connected World of 2020 with Deep Learning IoT*. Seattle, WA: Create Space Independent Publishing Platform, 2016.

Kastrenakes, J. "Nest Can Now Use Your Phone to Tell When You've Left the House." *The Verge*, March 10, 2016. **theverge.com/2016/3/10/11188888/nest-now-uses-location-for-home-away-states-launches-family-accounts** (accessed April 2018).

Khoury, A. "You Can Now Hail a Ride in a Fully Autonomous Vehicle, Courtesy of Waymo." *Digital Trends*, February 17, 2018.

Korosec, K. "Toyota Is Using Nvidia's Supercomputer to Bring Autonomous Driving to the Masses." *The Verge*, May 10, 2017.

Koufopoulos, J. "9 Examples of the Internet of Things That Aren't Nest." *Percolate*, January 23, 2015.

Kvitka, C. "Navigate the Internet of Things." January/February 2014. **oracle.com/technetwork/issue-archive/2014/14-jan/o14interview-utzschneider-2074127.html** (accessed April 2018).

Lacey, K. "Higher Ed Prepares for the Internet of Things." *University Business*, July 27, 2016. **universitybusiness.com/article/higher-prepares-internet-things** (accessed April 2018).

Libelium. "Smart Factory: Reducing Maintenance Costs and Ensuring Quality in the Manufacturing Process." *Libelium World*, March 2, 2015. **technology.ihs.com/531114/the-internet-of-everything-needs-a-fabric** (accessed April 2018).

Manyika, J., M. Chui, P. Bisson, J. Woetzel, R. Dobbs, J. Bughin, and D. Aharon. "Unlocking the Potential of the Internet of Things." *McKinsey Global Institute*, June 2015.

Marcus, J. "CNH Industrial Halves Product Downtime with IoT." *Product Lifecycle Report*, May 6, 2015.

Martin, E. "AI May Have Your Health and Finances on Record Before the Year Is Out." *FutureFive*, July 20, 2017. **futurefive.co.nz/story/five-ways-ai-machine-will-affect-your-life-and-business-year/** (accessed April 2018).

McCafferty, D. "How the Internet of Things Is Changing Everything." *Baseline*, June 16, 2015.

McGrath, J. "Haier Wants You to Live Smaller and Smarter with Its New Appliances." *Digital Trends*, January 5, 2016. **digitaltrends.com/home/haier-shows-off-u-smart-appliances-at-ces-2016** (accessed April 2018).

McLellan, C. "Internet of Things in the Enterprise: The State of Play." *ZDNet.com*, February 1, 2017a. **zdnet.com/article/enterprise-iot-in-2017-the-state-of-play/** (accessed April 2018).

McLellan, C. "Cybersecurity in an IoT and Mobile World." Special Report. *ZDNet*, June 1, 2017b.

Meola, A. "What Is the Internet of Things (IoT)? Meaning & Definition." *Business Insider*, May 10, 2018.

Miller, M. *The Internet of Things: How Smart TVs, Smart Cars, Smart Homes, and Smart Cities Are Changing the World*. Indianapolis, IN: Que Publishing, 2015.

Miller, R. "IoT Devices Could Be Next Customer Data Frontier." *TechCrunch*, March 30, 2018.

Morris, C. "Ordinary Home Appliances Are About to Get Really Sexy." *Fortune.com*, January 6, 2016. **fortune.com/2016/01/06/home-appliances-ces-2016** (accessed April 2018).

Morris, S., D. Griffin, and P. Gower. "Barclays Puts in Sensors to See Which Bankers Are Their Desks." *Bloomberg*, August 18, 2017.

Murray, M. "Intel Lays Out Its Vision for a Fully Connected World." *PC Magazine*, August 16, 2016.

Ohnsman, A. "Our Driverless Future Begins as Waymo Transitions to Robot-Only Chauffeurs." *Forbes*, November 7, 2017.

Park, H. "The Connected Customer: The Why Behind the Internet of Things." *Blue Hill Research*. White Paper. January 2017.

Perkins, E. "Securing the Internet of Things." Report G00300281. *Gartner Inc.*, May 12, 2016.

Pitsker, K. "Put Smart Home Technologies to Work for You." *Kiplinger's Personal Finance*, October 2017.

PTC, Inc. "Internal Transformation for IoT Business Model Reshapes Connected Industrial Vehicles." *PTC Transformational Case Study*, November 12, 2015. **ptc.com/~/media/Files/PDFs/IoT/J6081_CNH_Industrial_Case_Study_Final_11-12-15.pdf?la=e** (accessed April 2018).

Pujari, A. "Becoming a Smarter Manufacturer: How IoT Revolutionizes the Factory." *Enterprise Innovation*, June 5, 2017.

Rainie, L., and J. Anderson. "The Internet of Things Connectivity Binge: What Are the Implications?" *PewInternet.com*, June 6, 2017.

SAS. "SAS Analytics for IoT: Smart Cities." *SAS White Paper 108482_G14942*, September 2016.

SAS. "5 Steps for Turning Industrial IoT Data into a Competitive Advantage." *SAS White Paper 108670_G456z 0117.pdf*, January 2017.

Scannell, B. "High Performance Inertial Sensors Propelling the Internet of Moving Things." Technical Article. *Analog Devices*, 2017.

Schwartz, S. *Street Smart: The Rise of Cities and the Fall of Cars*. Kindle Edition. New York, NY: Public Affairs, 2015.

Shah, S. "HPE, Tata to Build 'World's Largest' IoT Network in India." *Internet of Business*, February 27, 2017. **internetofbusiness.com/hpe-tata-largest-iot-network-india/** (accessed April 2018).

Sharda, R., et al. *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. 4th ed. New York, NY: Pearson, 2018.

Sinclair, B. *IoT Inc.: How Your Company Can Use the Internet of Things to Win in the Outcome Economy*. Kindle Edition, New York, NY: McGraw-Hill Education, 2017.

Solomon, S. "Israel Smart-Roads Startup Nabs Prestigious EY Journey Prize." *The Times of Israel*, October 26, 2017.

Sorkin, A. "Larry Page's Flying Taxis Now Exiting Stealth Mode." *The New York Times*, March 12, 2018.

Staff. "Study Reveals Dramatic Increase in Capabilities for IoT Services." *Information Management*, May 5, 2017.

Technavio. "Smart Sensors for the Fourth Industrial Revolution: Molding the Future of Smart Industry with Advanced Technology." *Technavio.com*, September 12, 2017.

Tokuoka, D. *Emerging Technologies: Autonomous Cars*. Raleigh, NC: Lulu.com, 2016.

Tomás, J. "Smart Factory Tech Defining the Future of Production Processes." *RCR Wireless News*, March 28, 2016.

Townsend, A. *Smart Cities: Big Data, Civic, Hackers and the Quest for a New Utopia*. New York, NY: W. W. Norton, 2013.

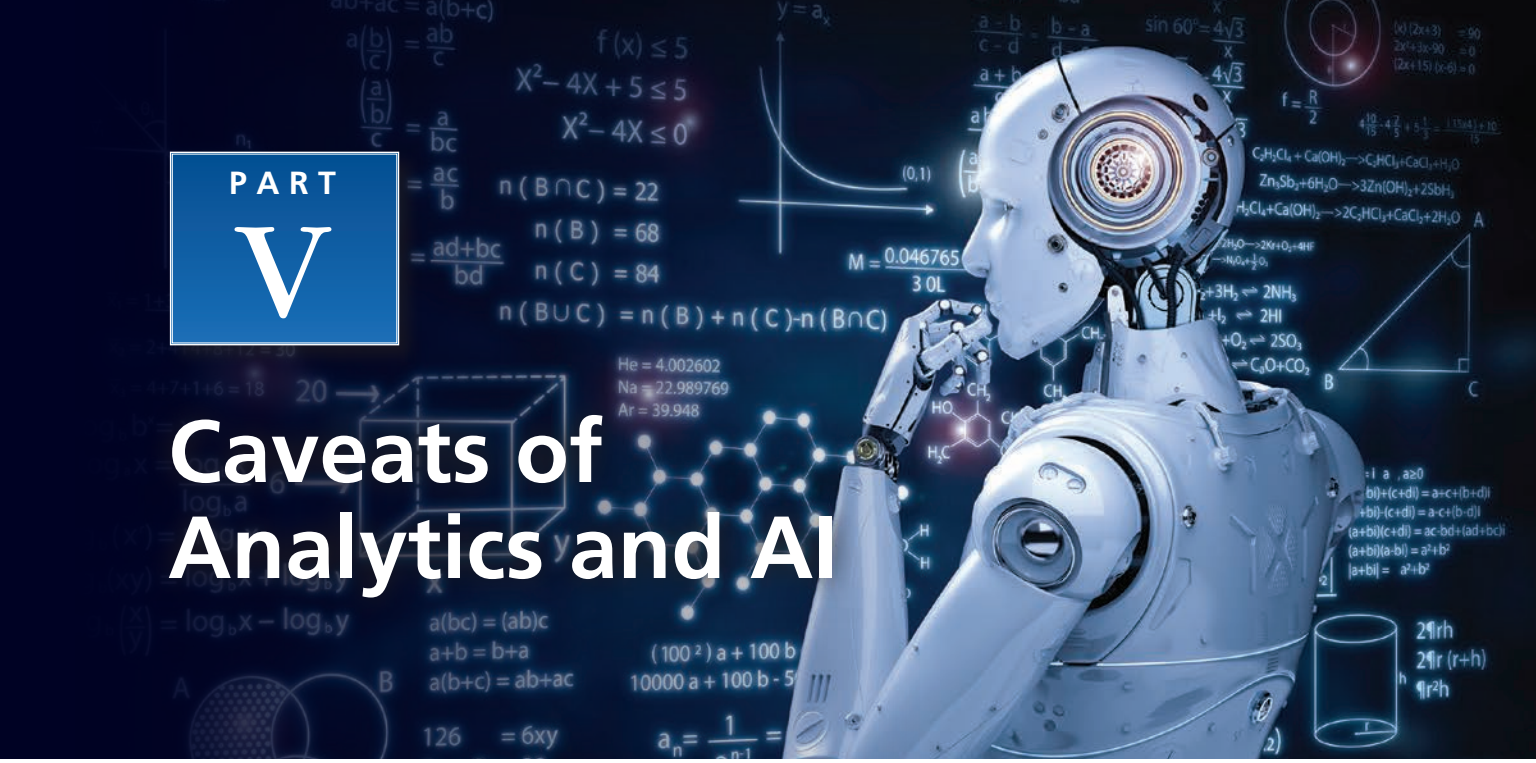
Twentyman, J. "Athens International Airport Turns to IoT for Environmental Monitoring." *Internet of Business*, September 4, 2017.

Venkatakrishnan, K. "Are Connected Consumers Driving Smart Homes?" *Enterprise Innovation*, May 31, 2017.

Violino, B. "19 Top Paying Internet-of-Things Jobs." *Information Management*, October 25, 2017.

Weinreich, A. "The Future of the Smart Home: Amazon, Walmart, & the Home That Shops for Itself." *Forbes*, February 1, 2018.

Weldon, D. "Steps for Getting an IoT Implementation Right." *Information Management*, October 30, 2015.



PART

V

Caveats of Analytics and AI

CHAPTER 14

Implementation Issues: From Ethics and Privacy to Organizational and Societal Impacts

LEARNING OBJECTIVES

- Describe the major implementation issues of intelligent technologies
- Discuss legal, privacy and ethical issues
- Understand the deployment issues of intelligent systems
- Describe the major impacts on organizations and society
- Discuss and debate the impacts on jobs and work
- Discuss the arguments of utopia and dystopia in a debate of the future of robots and artificial intelligence (AI)
- Discuss the potential danger of mathematical models in analytics
- Describe the major influencing technology trends
- Describe the highlights of the future of intelligent systems

In this concluding chapter, we cover a variety of issues related to the implementation and future of intelligent systems. We begin our coverage with technological issues such as security and connectivity. Then, we move to managerial issues that cover legality, privacy, and ethics. We next explore the impacts on organizations, society, work and jobs. Then, we review technology trends that point to the future.

Note. In this chapter we refer to all technologies covered in this book as *intelligent technologies* or *intelligent systems*.

This chapter has the following sections:

14.1 OPENING VIGNETTE: Why Did Uber Pay $245 Million to Waymo?

In early 2018, Uber Technologies, Inc. paid $245 million worth of its own shares to Waymo Self-Driving Cars (a subsidiary of Alphabet). The payment was made to settle a lawsuit filed by Waymo alleging that Uber was using Waymo's stolen proprietary technology.

THE BACKGROUND OF THE CASE

The lawsuit relates to the protection of intellectual property (trade secrets) owned by Waymo. As you may recall from Section 13.9, Waymo pioneered the self-driving car. A former engineer of Waymo (named Levandowski) allegedly illegally downloaded 14,000 of Waymo self-driving related confidential files. Worse than that, Levandowski may have convinced several top engineers of Waymo to leave Waymo and join him to create a start-up, Otto Company, for developing self-driving vehicles. Uber acquired Otto Company. For Uber, self-driving cars are essential for profitable growth when Uber will use such cars in a car-hailing system. Uber is a major car-hailing company that plans to move from sharing cars owned by individuals to the car-hailing business where self-driven cars will be owned by Uber and/or by car manufacturers. This way the profit for Uber could be much higher. Furthermore, Uber plans to operate driverless taxi fleets.

THE LEGAL DISPUTE

The legal dispute is very complicated. It deals with *intellectual property* and the ability of high technology employees to work after leaving their jobs for competitors.

Lawyers from Waymo claimed potentially huge damage if the Waymo trade secrets are used by the competitors. Waymo's legal team based their case on a digital-forensics investigation that proved that Levandowski deliberately copied the confidential files and then tried to cover this downloading. Note that Uber did not steal trade secrets, but hired Levandowski, who had these secrets.

From a legal point of view, the case was unique, being the first related to self-driving cars, so there were no previous cases to rely on. The two companies are large tech companies in Silicon Valley.

Employees that leave companies are interviewed and reminded that they signed an agreement regarding trade secrets they acquired when working for the company they leave. Levandowski said in his exit interview from Waymo that his future plans did not include competing activities that may compete with Waymo's self-driving cars. However, he had already met with Uber and sold it his new company, Otto Trucking. It became very clear that both Uber and Levandowski were not telling the truth.

WHY DID THEY SETTLE?

The rivals settled after four days in court. The case was in front of a jury, a fact that introduced an uncertain element to the case.

Waymo agreed to settle since, to win the case, it had to prove actual damage, which it was unable to do. Future damage is very difficult to compute. Furthermore, there was

no evidence that Uber was using any of Waymo's trade secrets, and Uber had already fired Levandowski.

Uber agreed to pay Waymo because the legal case constituted a possible delay in its development of self-driving cars, which is critical for the future of Uber. Also, legal fees were mounting (Uber is involved in several other legal issues related mostly to its drivers). Fighting Waymo did not ensure success given the deep pockets of Google. Actually, Waymo sent a clear message that it would protect its leading self-driving cars' position at any cost.

CONCLUSION

- Uber paid about one-third of 1 percent in shares of its company. Uber was valued at $70 billion (January 2018), which makes the payment equivalent to $245 million. Uber is planning on going public, which may increase its valuation.
- Uber agreed not to incorporate Waymo's confidential information into its existing or future technology. This was a major condition of Waymo.
- The reason why this dispute is important to both companies is that the autonomous vehicle market could be worth $7 trillion by 2050 (per Marshall and Davies, 2018). *Note:* It is equal to about one-third of the total current U.S. national debt.
- There has been a major emerging change in the nature of the self-driving cars' competition between 2016, when the legal dispute settled in July 2018.

Today, there are many more competitors and much more publicly known technologies and processes (i.e., fewer trade secrets). Finally, companies need to tell their employees what is not a trade secret.

➤ DISCUSSION QUESTIONS FOR THE OPENING VIGNETTE

1. Identify the legal issues involved in this case.
2. Why do you think Waymo agreed to take Uber's shares instead of money?
3. What is the meaning of intellectual property in this case?
4. The presiding federal judge said at the end: "This case is now ancient history." What did he mean to say?
5. Summarize the potential damages to the two parties if they had continued with the legal dispute.
6. Summarize the benefits of the settlement to both sides.

WHAT CAN WE LEARN FROM THIS VIGNETTE

Self-driving cars are a major product of intelligent systems and artificial intelligence (AI) with huge potential benefits to its participants. Also, inevitable is the strong competition in the industry and the importance of trade secrets acquired along the way. Legal disputes are common in competitive situations, and the protection of intellectual property is critical. Intellectual property protection is one topic presented in our concluding chapter. Other issues that are related to the implementation of intelligent systems and are discussed in this chapter are ethics, security, privacy, connectivity, integration, strategy, and top management roles.

We also learned in this vignette about the future importance of the new technology of autonomous vehicles. This technology may have a huge impact on organizations and their structure and operation. In addition, we discuss in this chapter the societal impacts of intelligent systems, and particularly their impact on work and jobs. We also explore some potential unintended consequences of intelligent systems. Finally, we explore the

potential future of intelligent systems and introduce the big debate regarding the dangers versus possible benefits of intelligent systems and particularly robots and AI.

Sources: Compiled from A. Marshall & A. Davies. (2018, February 9). "The End of Waymo v. Uber Marks a New Era for Self-Driving Cars: Reality." *Wired*; A. Sage, et al. (2018, February 9). "Waymo Accepts $245 Million and Uber's 'Regret' to Settle Self-Driving Car Dispute." *Reuters (Business News)*; K. Kokalitcheva. (2017, May 9). "The Full History of the Uber-Waymo Legal Fight." *Axios*.

14.2 IMPLEMENTING INTELLIGENT SYSTEMS: AN OVERVIEW

Now that you have learned the essentials of analytics, data science, artificial intelligence, and decision support activities, you may be tempted to ask: What can I do with all this in my organization? You learned about the great benefits and you read about numerous companies that use intelligent systems. So, what you should do next? First read some of the resources recommended in this book so you will have a better understanding about the technologies. Next, read this chapter that deals with the major issues that are involved in implementing intelligent systems in organizations.

Implementing business analytic/AI systems can be a complex undertaking. In addition to specific issues found in intelligent systems, there are issues that are common to many other computer-based information systems. In this section, we describe the major types of issues, some of which are discussed in this chapter. For several success AI implementation factors revealed in a survey of 3000 executives, see Bughin, McCarthy, and Chui (2017).

The Intelligent Systems Implementation Process

This chapter is divided into three parts. In the first part, we describe some managerial-related implementation issues. In the second part, we describe the impacts of intelligent technologies on organizations, management, work, and jobs. The last part deals with technology trends and the future of intelligent technologies.

The implementation process of intelligent systems is similar to the generic process of other information systems. Therefore, we will present it only briefly. The process is illustrated in Figure 14.1.

THE MAJOR STEPS OF IMPLEMENTATION The major steps are:

Step 1 Need assessment. Need assessment needs to provide the *business case* for the intelligent systems, including their major parts. (This is a generic IT step and will not be discussed here.)

Step 2 Preparations. In this step, it is necessary to examine the organization readiness for analytics and AI. It is necessary to check available resources, employees'

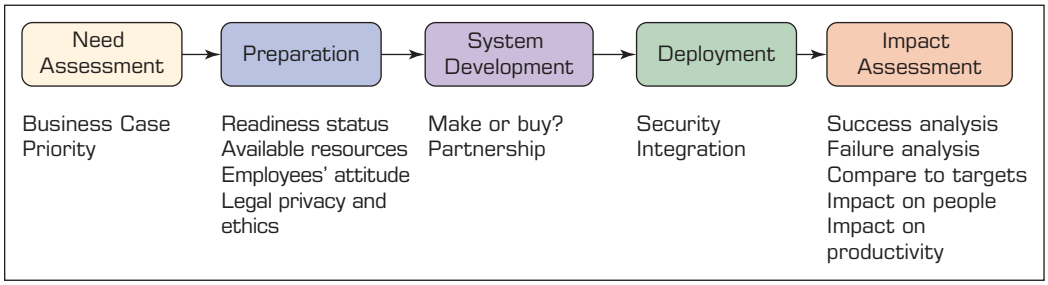


FIGURE 14.1 **Implementation Process.** Drawn by E. Turban

attitudes for the change, projects' priorities, and so on. This generic IT activity will not be discussed here. However, it is useful to think about legal, privacy, and ethical issues as they are related to intelligent technologies as described in Section 14.3.

Step 3 System acquisition. Organizations need to decide on in-house or outsourcing approach (make or buy) or on a combination of the two and possibly with partnership with a vendor or another company. A consultant may help at this step. It is a generic IT step that will not discussed here.

Step 4 System development. Regardless of who will develop the system, certain activities need to be done. These include security, integration with other systems, project management preparation, and other activities. Again, many of those are generic and will not be described here. Only selected ones are described in Section 14.4.

Step 5 Impact assessment. It is necessary to check the performance of the systems against plans. Again, this is a generic issue that will not be covered here.

The Impacts of Intelligent Systems

Intelligent systems are impacting all our lives and many businesses and other organizations. It is much easier to find what is not impacted than what is impacted. In this section, we divide these impacts into three categories as shown in Figure 14.2 with the section numbers where they are presented. We exclude from this list the impact on individuals and quality of life, which is a very large field (health, education, entertainment, crime fighting, social services, etc.).

Example

Here is the example in the entertainment field. In the near future, when you go Disneyland, Disney World, or one of the Disney International Parks, you will see high-flying acrobatic robots. You will see them everywhere there, and it is amazing. For a preview, watch the following videos: **money.cnn.com/video/news/2018/07/04/disney-robots-acrobatics-stuntronics-animatronics.cnnmoney/index.html** and **youtube.com/watch?v=Z_QGsNpI0J8**.

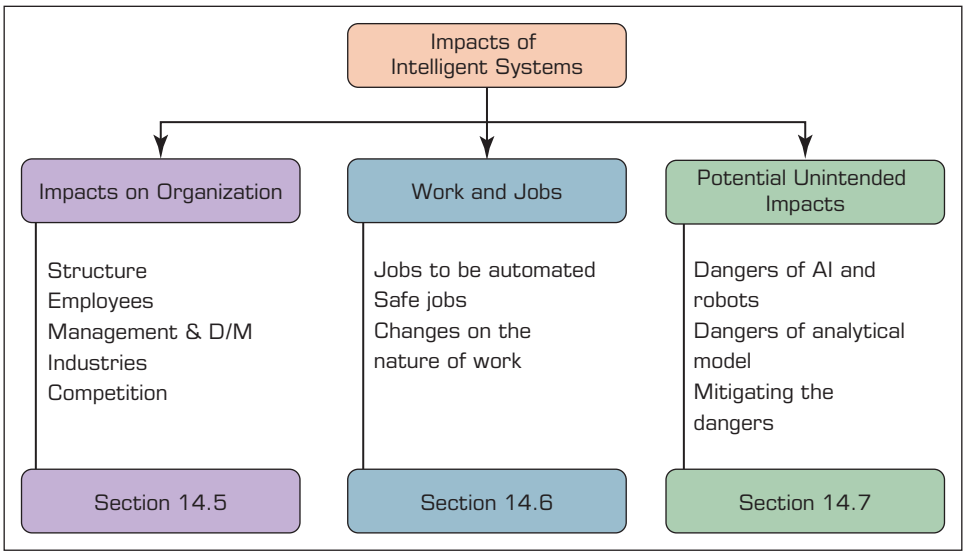


FIGURE 14.2 Impact Landscape. Drawn by E. Turban

SECTION 14.2 REVIEW QUESTIONS

1. List the major steps in the implementation process.
2. Why is implementation an important subject?
3. Describe the major impact areas of intelligent systems.

14.3 LEGAL, PRIVACY, AND ETHICAL ISSUES

As data science, analytics, cognitive computing, and AI grow in reach and pervasiveness, everyone may be affected by these applications. Just because something is doable through technology does not make it appropriate, legal, or ethical. Data science and AI professionals and managers have to be very aware of these concerns. Several important legal, privacy, and ethical issues are related to intelligent technologies and they are interrelated. For example, several privacy issues are parts of ethics or have legal aspects. Here we provide only representative examples and sources as pointed out in Chapter 1. Our goal here is only to give the reader an exposure to these issues. For why should we care about the legal, ethical, and privacy of AI, see Krigsman (2017).

Legal Issues

The introduction of intelligent technologies may compound a host of legal issues already relevant to computer systems. For example, questions concerning liability for the actions of advice provided by intelligent machines are beginning to be considered. In this section, we provide a sample of representative issues. Many more exist.

In addition to resolving disputes about the unexpected and possibly damaging results of some intelligent systems (see the opening vignette and Section 14.7), other complex issues may surface. For example, who is liable if an enterprise finds itself bankrupt as a result of using the advice of an AI-based application? Will the enterprise itself be held responsible for not testing the system adequately before entrusting it with sensitive or volatile issues? Will auditing and accounting firms share the liability for failing to apply adequate auditing tests? Will the software developers of intelligent systems be jointly liable? As self-driving cars become more common, who is liable for any damage or injury when a car's sensors, network, or AI system fail to function as planned? A recent case involving a Tesla car accident where the driver died in a crash while the car was allegedly on "autopilot" mode has brought this issue to the front pages of newspapers and the legal profession.

A SAMPLE OF AI POTENTIAL LEGAL ISSUES

- What is the value of an expert opinion in court when the expertise is encoded in a computer?
- Who is liable for wrong advice (or information) provided by an intelligent application? For example, what happens if a physician accepts an incorrect diagnosis made by a computer and performs a procedure that results in the death of a patient?
- What happens if a manager enters an incorrect judgment value into an intelligent application and the result is damage or a disaster?
- Who owns the knowledge in a knowledge base (e.g., the knowledge of a chatbot)?
- Can management force experts to contribute their expertise to an intelligent system? How will they be compensated?
  - Is it okay for self-driving cars with in-vehicle back-up drivers to drive on public roads? (Yes, in a few states, notably in California.)
  - Who should regulate driverless car: cities, states, or the federal government?
  - U.S. federal regulators are creating national laws for self-driving cars (for safe driving).

- Should delivery robots be allowed on sidewalks? (Not in San Francisco but in some European cities)
- Are drivers of Uber and similar companies self-employed? (Not in London, the United Kingdom)
- Should robots have human rights? (What if they are citizens like Sophia in Saudi Arabia?) If they get rights, should they have legal responsibilities as well?
- Should we legalize robot taxis? Would this make trips cheaper? (Yes in Singapore and other places, and it can be cheaper)

Source: Turban, Introduction to Information Technology, 2nd edition, John Wiley & Sons, 2006.

Example: Intellectual Property Protection

The opening vignette directed our attention to a legal issue that is very important for technology-related companies: the ownership and protection of intellectual property.

LEGAL ISSUES OF INTELLIGENT TECHNOLOGIES Several of the ethical issues described later need to be combined with legal issues. For example, take robots' legal rights. Do we need these rights? What for (an ethical issue)? Then, it will be necessary to develop the legal rights. Facebook, for example, has had legal issues regarding face recognition. Safety rules for robots were developed a long time ago. At the moment, there are very few laws regarding intelligent technologies. Most of the laws relate to safety.

AI AND LAW In addition to laws related to robotics and AI, there is a subfield of AI that is concerned with AI applications to the legal profession and the solution of some legal problems. According to Donahue (2018), the following are some major topics:

- Analyzing legal-related data (e.g., regulatory conflicts) to detect pattern
- Providing legal advice to consumers (e.g., see **DoNotPay.com**).
- Document review
- Analyzing contracts
- Supporting legal research
- Predicting results (e.g., likelihood to win)
- AI impact on the legal profession.

AI can execute routine legal-related tasks such as managing documents and drafting contracts. For details, see Kahn (2017). For 35 applications in law and legal practice see Rayo (2018). Legal issues may be strongly connected to our next topic, privacy.

Privacy Issues

Privacy means different things to different people. In general, **privacy** is the right to be left alone and the right to be free from unreasonable personal intrusions. Privacy has long been related to legal, ethical, and social issues in many countries. The right to privacy is recognized today in every state of the United States and by the federal government either by statute or by common law. The definition of *privacy* can be interpreted quite broadly. However, the following two rules have been followed fairly closely in past court decisions: (1) The right of privacy is not absolute. Privacy must be balanced against the needs of society. (2) The public's right to know is superior to the individual's right to privacy. These two rules show why it is difficult, in some cases, to determine and enforce privacy regulations. Privacy issues online have specific characteristics and policies. One area where privacy may be jeopardized is discussed next. Privacy issues are getting more

and more important as the amount of data generated on the Internet is increasing exponentially, and in many cases it is lightly secured. For an overview of privacy as it relates to AI, see Provazza (2017).

COLLECTING INFORMATION ABOUT INDIVIDUALS Intelligent technologies aim to provide targeted services and marketing to consumers; they do so by collecting information about these customers. In the past, the complexity of collecting, sorting, filing, and accessing information manually from numerous government agencies and other public databases was, in many cases, a built-in protection against the misuse of private information. The Internet in combination with large-scale databases has created an entirely new dimension of accessing and using data. The inherent power in intelligent systems that can access vast amounts of data and interpret them can be used for the good of society. For example, by analyzing records with the aid of business analysis, it is possible to eliminate or reduce fraud, crime, government mismanagement, tax evasion, welfare cheating, family-support filching, employment of illegal workers, and so on. However, what price must the individual pay in terms of loss of privacy so that the government can better apprehend criminals? The same is true on the corporate level. Private information about employees may aid in better corporate decision making, but the employees' privacy may be compromised.

The use of AI technologies in the administration and enforcement of laws and regulations may increase public concern regarding privacy of information. These fears, generated by the perceived abilities of AI, will have to be addressed at the outset of almost any AI development effort.

VIRTUAL PERSONAL ASSISTANTS Amazon's Echo/Alexa and similar devices listen to what is going on. They also may take photos. In other words, your voice assistant is spying on you.

Most advanced is the Echo/Alexa pair. You can ask Alexa to buy Amazon products today. Amazon and Google filed for a patent that will enable the virtual assistants in your home to advertise and sell you products. Privacy advocates are not happy, but customers may be. For example, Elgen (2017) describe how Alexa acts as a fashion consultant, using *style check*. The system combines the knowledge of a fashion specialist and AI knowledge. A recommendation provides you with two photos at a time, telling you which one to buy (based on color, current trends, etc.). To make it useful, Amazon is improving the privacy. This may not be easy since your record is stored in Amazon's cloud.

Huff (2017) provides arguments about the risks of the assistant and the protection provided by Amazon.

MOBILE USER PRIVACY Many users are unaware of the private information being tracked through their smartphone usage. Many apps collect user data that track each phone call as it moves from one cell tower to another, from GPS-enabled devices that transmit users' locations, and from phones transmitting information at Wi-Fi hotspots. Major app developers claim that they are extremely careful and protective of users' privacy, but it is interesting to note how much information is available through the use of a single device, especially when smartphones contain more and more AI components.

PRIVACY IN IOT NETWORKS For privacy and security of the Internet of Things (IoT), see Hu (2016). More data are flowing with IoT networks. Note that AI data privacy issues are on the rise, especially when AI deals with consumers' data. There is a growing amount of data collected, for example, by machine learning and chatbots. Also, in the enterprise, employers collect and analyze more data on employees. How do we protect the data and guard against their misuse?

RECENT TECHNOLOGY ISSUES IN PRIVACY AND ANALYTICS With the growth of Internet users in general and mobile device users in particular, many companies have started to employ intelligent technologies to develop profiles of users on the basis of their device usage, surfing, and contacts. *The Wall Street Journal* has an excellent collection of articles titled "What They Know" (**WallStreetJournal.com**, 2016). These articles are constantly updated to highlight the latest technology and privacy/ethical issues. One of the companies mentioned in this series is Rapleaf (now part of Towerdata). Rapleaf's technology claims to be able to provide a profile of a user just by knowing his or her e-mail address. Clearly, Rapleaf's technology enables it to gather significant related information. Another company that aims to identify devices on the basis of their usage is BlueCava, which recently merged with Qualia (**Qualia.com**). Qualia's BlueCava technology attaches a personal profile to be able to recognize a user as one individual or a household, even though the user may be working with multiple mobile devices and laptops. All of these companies employ analytics such as clustering and association mining to develop profiles of users. Of course, many of the analytics start-ups in this space claim to honor user privacy, but violations are often reported. For example, Rapleaf was collecting unauthorized information from Facebook users and was subsequently banned from Facebook. One user reported that an hour after he gave his e-mail address to a company that specializes in user information monitoring (**reputation.com**), the company was able to discover his Social Security number. So, violations of privacy create fears of criminal conduct regarding information. This area is a big concern overall and needs careful study. These examples not only illustrate the power of analytics in being able to learn more about target customers but also serve as a warning to AI and analytics professionals about being sensitive to privacy and ethical issues.

Another related application area of privacy concerns is analyzing employee behaviors on the basis of data collected from sensors that employees wear in a badge. One company, Humanyze, has reported several such applications of its sensor-embedded badges. These sensors track all movements of an employee.

Example: Using Sensors and IoT to Observe Bankers at Barclays Bank

Using heat and motion sensors, Barclays tracks how long its bankers are at their desks. The system was installed in the London, United Kingdom, branches. The formal explanation was to find out the occupancy of the cubes in the bank to optimally allocate and possibly reduce office space. The IoT network provided dashboards showing which workstations (cubes) were underutilized, and what the usage trend was. The bank informed the employees and the union that this project did not measure productivity, only space utilization. The results can be used to better manage energy consumption in the cubes and to schedule a flexible work environment. As a result, Barclays was able to save office space and rent it out for $45 million a year.

The bank uses a similar tracking system to find out how much time that different types of employees spend with customers. The union is watching this IoT application carefully to ensure that it is not used to spy on employees. Other banks in England use similar systems. For details, see Bloomberg News (2017).

Of course, situations like those described create major privacy concerns. Should companies be able to monitor their employees this intrusively?

Finally, there is a possibility of ransomware, or hackers' attacks on robots, which could be used against businesses whose employees use such robots. Smith (2018) reported on research that identified 50 vulnerabilities in robots. Ransomware attacks may interrupt operations, forcing organizations to pay substantial ransoms.

OTHER ISSUES OF POTENTIAL PRIVACY VIOLATION The following are some more examples of potential privacy violations in the intelligent technology world:

- Delaware police are using AI dashcoms to look for fugitives in passing cars. Photos and videos taken are sent to the clouds and analyzed there by AI algorithms.
- Facebook's face recognition systems create concerns regarding privacy protection.
- Epicenter offers its employees a microchip implant. It acts like a swipe card, opens doors, buys you food in the company store, and much more. But management can track you too. It is given only to volunteers.

Who Owns Our Private Data?

With the recent growth of data from our use of technology and the companies' ability to access and mine it, the privacy debate also leads to the obvious issue of whose property any user's data is; see Welch (2016) for highlights in this issue in a *Bloomberg Businessweek* column. Take an example of a relatively new car. The car is equipped with many sensors starting with tire pressure sensors to GPS trackers that can keep track of where you have gone, how fast you were driving, when you changed lanes, and so on. The car may even know the passenger's weight added to the front seat. As Welch notes, a car connected to the Internet (most new cars are!) can be a *privacy nightmare* for the owner or a data "gold mine" for whoever can possess or analyze these data. A major battle is brewing between automobile manufacturers and technology providers such as Apple (CarPlay) and Google (Android Auto) on who owns these data and who can access them. This is becoming more crucial because as cars become smarter and eventually self-driving, the driver/passenger in the car could be a highly targeted prospect for marketers' services. For example, Google's Waze app collects GPS data from millions of users to track traffic information and help users find the best routes; but it also displays pop-up ads on the users' screens. Yelp, Spotify, and other apps popularly used in cars have similar approaches.

The bottom line is that intelligent systems professionals and users must be aware of the legal and ethical issues involved in collecting information that may be privileged or protected. Privacy issues are considered in many cases as important components of ethics.

Ethics Issues

Several ethical issues are related to intelligent systems. Personal values constitute a major factor in the issue of ethical decision making. The study of ethical issues is complex because of their multidimensional nature. One story that upset many users (although it was not illegal) some time ago was Facebook's experiment to present different News Feeds to the users and monitor their emotional reactions as measured by replies, likes, sentiment analysis, and so on. Most companies, including technology companies, run user testing to identify the features most liked or disliked and fine-tune their product offerings accordingly. Because Facebook is so large, running this experiment without the users' informed consent was viewed as unethical. Indeed, Facebook acknowledged its error and instituted a more formal review through Internal Review Boards and other compliance mechanisms for future testing.

Morgan (2017) said that it is necessary to be at the foundations of what AI does for both vendors and customers in order to stay ethical and have transparency of each situation. This way people can stay honest and adhere to the goals of AI, so it can play a significant role in our life and work. For how ethical issues interfere with Alphabet's (Google) initiatives, see Kahn (2017).

Ethical Issues of Intelligent Systems

Many people have raised questions regarding ethical issues in AI, robotics, and other intelligent systems. For example, Bossmann (2016) raised the following issues:

1. What are their impact on jobs (see Section 14.5)?
2. How do machines (i.e., robots) affect our behavior and interactions?
3. How can wealth created by intelligent machines be distributed (e.g., Kaplan, 2016)?
4. How can intelligent applications mistakes be guarded against? For example, how long should training programs in machine learning be?
5. Can intelligent systems be fair and unbiased? How can bias in creation and operation of AI systems be eliminated?
6. How can intelligent applications be keep safe from adversaries?
7. How can systems be protected against unintended consequences (e.g., accidents in robot operations)? For example, Facebook researchers had to shut down an AI system that created its own poor language.
8. How can we stay in control of a complex intelligent system?
9. Should we develop robots' legal rights? How can we define and plan human treatment of intelligent machines?
10. Should we allow a self-governing robot society to exist with ours?
11. To what extent should we influence unintended robots' behavior (or even be able to)?
12. How would we get around the question of smart machine ownership?

Additional issues are:

- Electronic surveillance.
- Ethics in business intelligence (BI) and AI systems design.
- Software piracy.
- Invasion of individuals' privacy.
- Use of proprietary databases and knowledge bases.
- Use of personal intellectual property such as knowledge and expertise for the benefits of companies and the payment to the contributors.
- Accuracy of data, information, and knowledge.
- Protection of the rights of users.
- Accessibility to information by AI users.
- The amount of decision making to delegate to intelligent machines.
  - How AI can fail due to inappropriate ethics.
  - The ethics of legal analytics (Goldman, 2018).

Other Topics in Intelligent Systems Ethics

- Machine ethics is a part of the ethics of AI that is concerned with the moral behavior of artificially intelligent beings (per Wikipedia; see details there).
- Robotics is concerned with the moral behavior of designers, builders, and users of robots.
- Microsoft's Tay chatbot was closed due to its inability to understand many irrelevant and offending comments.
- Some are afraid that algorithm-based technologies, including AI, may become racists. We discuss this topic in Section 14.8. Also, see Clozel (2017).
- According to Spangler (2017), self-driving cars may one day face a decision of whom to save and whom to kill.
- Voice technologies enable the identification of callers to AI machines. This may be great on one hand, but it creates privacy concerns on the other.

- One area in which there are considerable ethical concerns (frequently combined with legal concerns) is the healthcare/medical field. Given the large efforts by Alphabet and IBM Watson initiatives, this is not surprising. For a discussion, see Bloomberg News (2017).

For comprehensive coverage of ethical issues in big data and data sharing, see Anon (2017). For principles for Big Data analysis, see Kassner (2017).

COMPUTER ETHICS IN GENERAL Computer ethics focuses on the behavior of people toward information systems and computers in general. The study of ethics in intelligent systems is strongly related to the ethics of computers and information systems in general. The following are some resources.

THE TEN COMMANDMENTS OF COMPUTER ETHICS This well-known document is published by cybercitizenship (**cybercitizenship.org/ethics/commandments.html**).

1. Thou shalt not use a computer to harm other people.
2. Thou shalt not interfere with other people's computer work.
3. Thou shalt not snoop around in other people's files.
4. Thou shalt not use a computer to steal.
5. Thou shalt not use a computer to bear false witness.
6. Thou shalt not use or copy software for which you have not paid.
7. Thou shalt not use other people's computer resources without authorization.
8. Thou shalt not appropriate other people's intellectual output.
9. Thou shalt not think about the social consequences of the program you write.
10. Thou shalt not use a computer in ways that show consideration and respect.

A major upcoming issue is that of ethics for autonomous vehicles. For example, who will develop them, how will they be programmed into the vehicles, and how will they be enforced? See Sharma (2017).

For review of ethical issue considerations in information research literature, see **nowpublishers.com/article/Details/ISY-012/**.

MIT Media Lab and the Center for Internet & Society at Harvard University manage an initiative to research ethical and governance topics in AI. SAS, a major analytical and AI vendor, proposed three essential steps for AI ethics as described in **sas.com/en_us/insights/articles/analytics/artificial-intelligence-ethics.html/**.

SECTION 14.3 REVIEW QUESTIONS

1. List some legal issues of intelligent systems.
2. Describe privacy concerns in intelligent systems.
3. In your view, who should own the data about your use of a car? Why?
4. List ethical issues in intelligent systems.
5. What are the 10 commandments of computer/information systems?

14.4 SUCCESSFUL DEPLOYMENT OF INTELLIGENT SYSTEMS

Many experts, consultants, and researchers provide suggestions regarding intelligent systems' successful deployment. Given the importance of the topic, it is clear that companies need to get ready for the mass arrival of AI and other intelligent technologies. Here are some topics related to deployment strategy:

- When to embark on intelligent projects and how to prioritize them.
- How to decide whether to do it yourself or use partners, or to outsource.

- How to justify investments in intelligent projects.
- How to overcome employees' resistance (e.g., fear of job loss).
- How to arrange appropriate people-robot teams.
- How to determine which decisions to fully automate by AI.
- How to protect intelligent systems (security) and how to protect privacy.
- How to handle possible loss of jobs and retraining of employees (Section 14.5).
- How to determine whether you have the necessary up-to-date technology.
- How to decide what support top management should provide.
- How to integrate the system with business processes.
- How to find qualified personnel for building and using intelligent systems.

For more strategy issues, see Kiron (2017). We cover only several topics in this section and provide references to more. Most of the implementation topics are generic in nature and will not be covered here.

Top Management and Implementation

According to Chui et al. (2017), from McKinsey & Company, "Senior executives need to understand the tactical as well as the strategic opportunities (of AI), redesign their organizations, and commit to helping shape and debate about the future of work." Specifically, the executives need to plan for integrating intelligent systems into their workplace, making a commitment to conduct a participating environment for the changes and provide sufficient resources. Snyder (2017) claims that many executives know that intelligent systems will change their business, but they do not do much about it.

KPMG, a large management service consultant, provides the following steps regarding digital labor:

> "KPMG's holistic approach-from strategy through execution will assist companies on each step of implementation. The steps are:
>
> - Establishing priority areas for technological innovation.
> - Developing a strategy and a plan for the employees.
> - Identify providers and partners for plans' execution.
> - Establishing a strategy and plans to realize benefits from the digital labor initiatives."

Source: KPMG Internal Audit: Top 10 in 2018, Considerations for impactful internal audit departments, © 2018 KPMG LLP.

A complete guide for KPMG is provided by Kiron (2017). It includes robotic process automation, enhanced process automation, and cognitive automation. For issues regarding leadership in implementation, see Ainsworth (2017).

System Development Implementation Issues

Since *AI* and *business analytics* are broad terms, describing several technologies whose maturity levels vary, implementation issues may vary considerably. Shchutskaya (2017) cites the following three major problems:

1. *Development approach.* Business analytic and AI systems require an approach different from that of other IT/computer systems. Specifically, it is necessary to identify and deal with different and frequently large data sources (see the opening vignette to Chapters 1 and 2). It is necessary to cleanse and curate these data. Also, if learning is involved, one needs to use machine training. Thus, special methodologies are needed.
2. *Learning from data.* Many AI and business analytics involve learning. The quality of the input data determines the quality of the applications. Also, the learning mechanism is important. Therefore, data accuracy is critical. In learning, systems must be

able to deal with changing environmental conditions. Data should be organized in databases, not in files.

3. *No clear view is available of how insights are generated.* AI, IoT, and business analytic systems generate insights, conclusions, and recommendations based on the analysis of the data collected. Given that data are frequently collected by sensors and there are different types of them, we may not have a clear view of the insights that are generated.

Related important areas include problems with Big Data, ineffective information access, and limited integration capabilities (discussed next).

Connectivity and Integration

As part of the development process, it is necessary to connect the AI and analytic applications to existing IT systems, including the Internet, and other intelligent systems.

Example

The Australian government commissioned Microsoft in August 2017 to build hyperscale cloud regions to unlock the power of intelligent technologies. The system is expected to dramatically modernize how the government processes data and delivers services to its citizens. The system can handle both unclassified and protected data. The infrastructure is built inside, or near, the government data centers. The system will enable the government to use innovative applications based on machine learning, bots, and language translation, and it will improve healthcare, education, social services, and other government operations. Finally, the system will increase both security and privacy protection.

Integration needs to be done with almost every system that is being impacted by AI or business analytic. For example, it is necessary to integrate intelligent applications both to a digital marketing strategy and to marketing implementation. For a discussion, see **searchenginejournal.com/artificial-intelligence-marketing/200852/**.

To overcome the integration difficulty, Huawei of China (a cellphone producer) is installing an AI system with its knowledge base inside the chips of its products. Other phones' manufacturers rely on connecting to the "cloud" to interact there with AI knowledge. For the implications on IoT connectivity, see Rainie and Anderson (2017). For considerations regarding IoT connectivity providers, see Baroudy et al. (2018).

Security Protection

Many intelligent applications are managed and updated in the "cloud" and/or connected to the regular Internet. Unfortunately, by adding Internet connection, new vulnerabilities may be created. Hackers use intelligent technologies to identify these vulnerabilities. For how criminals use AI and related issues, see Crosman (2017). In Section 14.7, we discuss the potential dangers of robotics. The safety of passengers in self-driving cars and others who may be involved in collisions with the self-driving cars is an important safety issue as well. Also, the safety of people working near robots has been researched for many decades. In addition, hacking robots, chatbots, and other intelligent systems are areas that require attention. Finally, the safety of robots themselves when they work on the streets is an issue. Some people attack them (see McFarland, 2017a and the video there).

Leveraging Intelligent Systems in Business

There are many ways to leverage intelligent systems, depending on the nature of the applications. Catliff (2017) suggests the following ways to do this, leveraging the

intelligent technology capabilities to increase efficiency and provide more customer care. Specifically, he suggested:

1. Customize the customer experience (e.g., for interactions with customers).
2. Increase customer engagement (e.g., via chatbots).
3. Use intelligent technologies to detect problems and anomalies in data.

Singh (2017a) recommends the following as critical success factors: discover, predict, justify, and learn from experience. Ross (2017) raised the issue of the need to upgrade employees' skills and build an empowered AI-savvy workforce. One of the most important issues is how to handle the fear of job loss of employees. This is discussed in Section 14.6.

Intelligent System Adoption

Most of the issues related to intelligent systems' adoption are the same as or similar to that of any information systems. For example, employees may resist change, management may not provide sufficient resources, there could be a lack of planning and coordination, and so on. To deal with such issues, Morgan Stanley drew ideas from hundreds of conversations with experts (see DiCamillo, 2018). One important issue is to have an appropriate deployment and adoption strategy that should work in harmony with the implemented technologies and the people involved. In general, the generic adoption approach to information systems should work here, too.

SECTION 14.4 REVIEW QUESTIONS

1. Describe the systems deployment process.
2. Discuss the role of top management in deploying intelligent systems.
3. Why is connectivity such an important issue?
4. Describe system development issues.
5. Discuss the importance of security and safety, and how to protect them.
6. Describe some issues in intelligent systems adoption.

14.5 IMPACTS OF INTELLIGENT SYSTEMS ON ORGANIZATIONS

Intelligent systems are important components in the information and knowledge revolution. Unlike the slower revolutions of the past, such as the Industrial Revolution, this revolution is taking place very rapidly and affecting every facet of our work and lives. Inherent in this transformation is the impact on organizations, industries, and managers, some of which are described in this section.

Separating the impact of intelligent systems from that of other computerized systems is a difficult task, especially because of the trend toward integrating, or even embedding, intelligent systems with other computer-based information systems. Intelligent systems can have both micro- and macro implications. Such systems can affect particular individuals and jobs as well as the work and structures of departments and units within an organization. They can also have significant long-term effects on total organizational structures, entire industries, communities, and society as a whole (i.e., regarding macro impact, see Sections 14.6 and 14.7).

Explosive growth in analytics, AI, and cognitive computing is going to have a major impact on the future of organizations. The impact of computers and intelligent systems can be divided into three general categories: organizational, individual, and societal. In each of these, computers may have many possible impacts. We cannot possibly consider all of them in this book, so in the next paragraphs we cover topics we feel are most relevant to intelligent systems and organizations.

New Organizational Units and Their Management

One change in organizational structure is the possibility of creating an analytics department, a BI department, a data science department, and/or an AI department in which analytics plays a major role. Such special units (of any type) can be combined with or replace a quantitative analysis unit, or it can be a completely new entity. Some large corporations have separate decision support units or departments. For example, many major banks have such departments in their financial services divisions. Many companies have small data science or BI/data warehouse units. These types of departments are usually involved in training in addition to consulting and application development activities. Others have empowered a chief technology officer over BI, intelligent systems, and e-commerce applications. Companies such as Target and Walmart have major investments in such units, which are constantly analyzing their data to determine the efficiency of marketing and supply chain management by understanding their customer and supplier interactions. On the other hand, many companies are embedding analytics/data science specialties within functional areas such as marketing, finance, and operations. In general, this is one area where considerable job opportunities currently exist. For a discussion of the need for a chief data officer, see Weldon (2018). Also, the need for a chief AI officer is discussed by Lawson (2017).

Growth of the BI and analytics has resulted in the formation of new units within IT companies as well. For example, a few years ago, IBM formed a new business unit focused on analytics. This group includes units in BI, optimization models, data mining, and business performance. More importantly, the group is focused not just on software but also significantly more on services/consulting.

Transforming Businesses and Increasing Competitive Advantage

One of the major impacts of intelligent systems is the transformation of businesses to digital ones. While such transformation has been going on with other information technologies for years, it has accelerated with intelligent technologies, mostly with AI.

In many cases, AI is only a supportive tool for humans. However, as AI has become more capable, machines have been able to perform more tasks by themselves or with people. The fact is that AI already is transforming some businesses. As seen in Chapter 2, AI already is changing all business functional areas, especially marketing and finance. The impact ranges from full automation of many tasks, including managerial ones, to an increase in human-machine collaboration (Chapter 11). A comprehensive description of how AI is driving digital transformation is provided by Daugherty and Wilson (2018), who concluded that businesses that will miss the AI-driven transformation would be in a competitive disadvantage. Batra et al. (2018) point to a similar phenomenon and urge companies to use AI and utilize it for a wave of innovations. For more on this topic, see Uzialko (2017).

USING INTELLIGENT SYSTEMS TO GAIN COMPETITIVE ADVANTAGE Use of intelligent technologies, and especially AI, is evidenced in many cases. For example, using robots, **Amazon.com** enabled the company to reduce cost and control online commerce. In general, by cutting costs, increasing customer experiences, improving quality, and speeding deliveries, companies will gain competitive advantage. Rikert (2017) describes conversations with CEOs about how AI and machine learning can beat the competitors. Andronic (2017) points to competitive advantage. The benefits include generating more demand (see Chapter 2), automating sales (Chapter 2), and identifying sales opportunities.

An important recent factor is the fact that new companies and blurring sector borders are influencing the competitive picture of many industries. For example, autonomous vehicles will impact the competition in the automotive industry.

According to Weldon (2017c), a smart use of analytics offers top competitive advantage. The author provides advice on how organizations can get the full benefits from analytics. An example of how **1-800-Flowers.com** is using analytics, AI, and other intelligent technologies to gain a competitive advantage is provided in Application Case 14.1.

Application Case 14.1 How 1-800-Flowers.com Uses Intelligent Systems for Competitive Advantage

1-800-Flowers.com is a leading online retailer of flowers and gifts. The company moved from telephone to online ordering in the mid-1990s. Since then, it has grown to over $1 billion in revenue and over 4,000 employees, despite fierce competition. In a world dominated by online giants such as **Amazon.com** and **Walmart.com**, and hundreds of other companies that sell online flowers and gifts, survival is not easy.

The company is using the following three key strategies:

- Enhancing the customer experience.
- Driving demand more efficiently.
- Building a workforce that supports the products and technology innovation (culture of innovation).

The company has been using intelligent technologies extensively to build a superb supply chain and to facilitate collaboration. Lately, it started to use intelligent systems to enhance its competitive strategies. Here are several technologies covered in this book that the company uses.

1. *Optimal customers experience.* Using SAS Marketing Automation and Data Management products, the retailer collects information regarding customers' needs and analyzes it. This information enables senders of flowers and gifts to find perfect gifts for any occasion. Senders want to make recipients happy, so appropriate recommendations are critical. The company uses advanced analytics and data mining from SAS to anticipate customers' needs. **1-800-Flowers.com** marketers can then communicate with customers more effectively. Using the newest tools, company data scientists and marketing analysts mine data more efficiently. Today customer expectations are higher than ever because it is much easier for customers to compare vendors' offerings online. Analytics and AI enable the company to understand its customers' sentiments. Now the company is able to understand the emotional reasoning behavior for purchasing decisions and customer loyalty. This change results in product recommendations described later.
2. *Chatbots.* **1-800-Flowers.com** has a bot on Facebook Messenger. As described in Chapter 12, such a bot can be useful as a source of information and as a vehicle for conversation. The company also offers chat on its Web site online, and chat using voice. In addition, mobile shoppers can use Google Assistant for voice ordering. The company also offers voice-enabled Alexa with its "one-shot intent" to expedite ordering.
3. *Customer service.* The company offers a portal and one-stop shopping similar to what **Amazon.com** offers, and self-service payment is available. The same capability is available when shopping with the company's bot on Facebook Messenger. Customers do not have to leave Facebook to complete an order.
4. *AI-based recommendation.* As you may recall from Chapter 12, e-commerce retailers excel by providing product recommendation (e.g., Amazon, Netflix). **1-800-Flowers.com** is doing the same thing, offering recommendation and advice on gifts from their brand's websites (e.g., Harry and David). The recommendations are generated by IBM's Watson and are offered as a "cognitive concierge," making online shopping feel as having an in-store experience. This AI-based service is known as GWYN (Gifts When You Need) at 1-800-flowers. Watson's natural language processing (NLP) enables easy shopper-machine conversations.
5. *Personalization.* SAS advanced analytics enables the company's marketing department to segment customers into groups with similar characteristics. Then the company can send

promotions targeted to the profile of each segment. In addition to e-mails, special campaigns are arranged. Based on the feedback, the company can plan and revise marketing strategy. SAS also helps the company to analyze the "likes" and "dislikes" of the customers. All-in-all, the intelligent systems help the company and its customers to make informed decisions.

Questions for Case 14.1

1. Why it is necessary to provide better customer experience today?
2. Why do data need sophisticated analytical tools?
3. Read the "Key benefit of SAS Marketing Automation." Which benefits do you think are used by **1-800-Flowers.com** and why?
4. Relate IBM Watson to "personalization."
5. Relate 'SAS Advanced Analytics' capabilities to their use in this case.
6. 'SAS Enterprise Miner' is used to do data mining. Explain what is done and how.
7. SAS has a product called 'Enterprise Guide' that **1-800-Flowers.com** uses. Find how it is used based on the tools' capabilities.

Sources: Compiled from J. Keenan. (2018, February 13). "1-800-Flowers.com Using Technology to Win Customers' Hearts This Valentine's Day." *Total Retail*; S. Gaudin. (2016, October 26). "1-800-Flowers Wants to Transform Its Business with A.I." *Computer World*; SAS. (n.d.). "Customer Loyalty Blossoms with Analytics." *SAS Publication*, **sas.com/en_us/customers/1-800-flowers.html/** (accessed July 2018).

Redesign of an Organization Through the Use of Analytics

An emerging area of research and practice is employing data science technologies for studying organizational dynamics, personnel behavior, and redesigning the organization to better achieve its goals. Indeed, such analytics applications are known as *People Analytics*. For example, analytics are used by HR departments to identify ideal candidates from the pool that submits resumes to the organization or even from broader pools such as LinkedIn. Note that with AI and analytics, managers will be able to have a larger span of control due, for example, to the advice managers and employees can get from virtual assistants. The increased span of control could result in flatter organizational structures. Also, managers' job descriptions may have to change.

A more interesting and recent application area relates to understanding employee behavior by monitoring their movements within the organization and using that information to redesign the layout or teams to achieve better performance. A company called Humanyze (previously known as Sociometric Solutions) has badges that include a GPS and a sensor. When employees wear these badges, all of their movement is recorded. Humanyze has reportedly been able to assist companies in predicting which types of employees are likely to stay with the company or leave on the basis of their interactions with other employees. For example, those employees who stay in their own cubicles are less likely to progress up the corporate ladder than those who move about and interact with other employees extensively. Similar data collection and analysis have helped other companies determine the size of conference rooms needed or even the office layout to maximize efficiency. According to Humanyze's Web site, one company wanted to better understand characteristics of its leaders. By analyzing the data from these badges, the company was able to recognize that the successful leaders indeed have larger networks with which they interact, spend more time interacting with others, and are also physically active. The information gathered across team leaders was used to redesign the work space and help improve other leaders' performance. Clearly, this may raise *privacy issues*, but within an organization, such studies may be acceptable. Humanyze's Web site has several other interesting case studies that offer examples of how Big Data technologies can be used to develop more efficient *team structures* and *organizational design*.

Intelligent Systems' Impact on Managers' Activities, Performance, and Job Satisfaction

Although many jobs may be substantially enriched by intelligent technologies, other jobs may become more routine and less satisfying. Some claim that computer-based information systems in general may reduce managerial discretion in decision making and lead managers to be dissatisfied. However, studies of automated decision systems found that employees using such systems, especially those who are empowered by the systems, were more satisfied with their jobs. If using an AI system can do routine and mundane work, then it should free managers and knowledge workers to do more challenging tasks.

The most important task of managers is making decisions. Intelligent technologies can change the manner in which many decisions are made and can consequently change managers' job responsibilities. For example, some researchers found that a decision support system improved the performance of both existing and new managers as well as other employees. It helped managers gain more knowledge, experience, and expertise and consequently enhanced the quality of their decision making. Many managers report that intelligent systems have finally given them time to get out of the office and into the field. They have also found that they can spend more time planning activities instead of putting out fires because they can be alerted to potential problems well in advance thanks to intelligent system technologies (see the opening vignette, Chapter 1).

Another aspect of the managerial challenge lies in the ability of intelligent technologies to support the decision-making process in general and strategic planning and control decisions in particular. Intelligent systems could change the decision-making process and even decision-making styles. For example, information gathering for decision making is completed much more quickly when algorithms are in use. Research indicates that most managers tend to work on a large number of problems simultaneously, moving from one to another as they wait for more information on their current problem. Intelligent technologies tend to reduce the time required to complete tasks in the decision-making process and eliminate some of the nonproductive waiting time by providing knowledge and information.

The following are some potential impacts of intelligent system on managers' jobs:

- Less expertise (experience) is required for making many decisions.
- Faster decision making is possible because of the availability of information and the automation of some phases in the decision-making process (see Chapters 2 and 11).
- Less reliance on experts and analysts is required to provide support to top managers and executives. Today, they can decide by themselves with the help of intelligent systems.
- Power is being redistributed among managers. (The more information and analysis capability they possess, the more power they have.)
- Support for complex decisions makes solutions faster to develop and of better quality.
- Information needed for high-level decision making is expedited or even self-generated.
- Automation of routine decisions or phases in the decision-making process (e.g., for frontline decision making and using automated decision making) may eliminate some managers.

Source: Decision Support And Business Intelligence Systems, Pearson Education India, 2008.

In general, it has been found that the job of middle managers is the most likely job to be automated. Midlevel managers make fairly routine decisions, which can be fully automated. Managers at lower levels do not spend much time on decision making.

Instead, they supervise, train, and motivate nonmanagers. Some of their routine decisions, such as scheduling, can be automated; other decisions that involve cognitive aspects may not be automated. However, even if managers' decisional role is completely automated, many of their other activities could not be automated or could only be partially automated.

Impact on Decision Making

Throughout the book, we illustrate how intelligent technologies improve or automate decision making. These technologies, of course, will impact managers' job. One aspect is the impact of intelligent technologies supported by the "cloud." An example is illustrated in Chapter 9, Figure 9.12. It illustrates the flow of data from data sources and services via an information service to analytical services for different types of decision making supported by analytics.

Uzialko (2017) describes how humans can use AI to predict and analyze the consequences of different potential solutions, streamlining the decision-making process. Also, by using machine learning and deep learning, more decisions can be automated.

One impact of intelligent systems is to support real-time decision making. A popular tool for doing just that is SAS® Decision Manager, which is described in Technology Insights 14.1.

TECHNOLOGY INSIGHT 14.1 SAS Decision Manager

SAS Real-Time Decision Manager (RTDM) is an analytics-based integrated product that is designed to support real-time decision making, which is necessary for helping companies respond to rapidly changing marketing, customers' demands, technology, and other business environments.

SAS answers the following questions:

1. **What does SAS** RTDM **do?** It combines SAS analytics with business logic and contact strategies to deliver enhanced real-time recommendations and decisions to interactive customer channels, such as Web sites, call centers, point of sales (POS) locations, and automated teller machines (ATMs).
2. **Why is SAS** RTDM **important?** It helps you make smarter decisions by automating and applying analytics to the decision process during real-time customer interactions. By successfully meeting each customer's specific needs at the right time, the right place, and in the right context, your business can become more profitable.
3. **For whom is SAS®** RTDM **designed?** It provides distinct capabilities for marketers who define communication strategies, executives who need reports on marketing effectiveness, business analysts who model and predict customer behavior, and campaign managers who create target customer segments.

The following are the key benefits of RTDM:

- Makes the right decisions every time, all the time.
- Realizes customer needs with the right offer, at the right time, in the right channel.
- Better allocates valuable IT resources.

The key features according to SAS Inc. are:

- Real-time analytics.
- Rapid decision process construction.
- Enterprise data throughout.
- Campaign testing.
- Automated self-learning analytical process.
- Connectivity.

For the details, visit "SAS Real-Time Decision Manager" and read the text there. Also you can download a white paper about RTDM there.

Discussion Questions

1. What improvements to the decision-making process are made by SAS RTDM?
2. What SAS products are embedded or connected to RTDM? (You need to read the Web site's details.)
3. Relate the product to product recommendation capability.

Source: SAS® Real-Time Decision Manager Make context-based marketing decisions during your real-time customer interactions. Copyright © 2018 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Used with permission.

Industrial Restructuring

A few authors have begun to speculate on the impact of AI, analytics, and cognitive computing on the future of industry. A few interesting resources to consult are Autor (2016), Ransbotham (2016), a special report by *The Economist* (Standage, 2016), and a book by Brynjolfsson and McAfee (2016). The report by *The Economist* is quite comprehensive and considers many dimensions of the impact of the current developments on industry and society. The main arguments are that technology is now enabling more and more tasks that were done by humans using computers. Automating work, of course, has happened before, since the time of the Industrial Revolution. What makes the change this time around significantly more far reaching is that the technology is enabling many cognitive tasks to be done by machines. And the speed of change is so radical that the likely impact on organizations and society will be very significant and at times unpredictable. These authors do not agree in their predictions, of course. Let us focus first on the organizational impacts. Ransbotham (2016) argues that cognitive computing will convert many jobs done by humans to be done by computers, thus reducing costs for organizations. The quality of output may increase as well in cognitive work, which has been shown in several studies that compare a human's performance with a machine. Everyone is aware of IBM Watson having won in *Jeopardy!* or Google's system winning in the game of GO against human champions. But many other studies in specific domains such as speech recognition and medical image interpretation have also shown similar superiority of automated systems when the task is highly specialized yet routine or repetitive. Also, because machines tend to be available at all hours and at all locations, an organization's reach may increase, resulting in easier scaling and thus greater *competition* among organizations. These organizational impacts mean that yesterday's top organizations may not remain at the top forever because cognitive computing and automation can challenge established players. This is the case in the automotive industry. Although traditional car companies are trying quickly to catch up, Google, Tesla, and other technology companies are disrupting industry structure by challenging the leaders of the automotive age. Analytics and AI are empowering many of these changes.

➧ SECTION 14.5 REVIEW QUESTIONS

1. List the impacts of intelligent systems on managerial tasks.
2. Describe new organizational units that are created because of intelligent systems.
3. Identify examples of analytics and AI applications used to redesign workspace or team behavior.
4. How is cognitive computing affecting industry structure and competition?
5. Describe the impacts of intelligent systems on competition.
6. Discuss the impact of intelligent systems on decision making.

14.6 IMPACTS ON JOBS AND WORK

One of the most discussed and debated topics in considering the impacts of intelligent systems is on jobs and work. There is a general agreement that:

- Intelligent systems will create many new jobs as automation always has.
- There will be a need to retrain many people.
- The nature of work will be changed.

The discussions, debates, and disagreements relating to the issues of when, how much, and how to deal with these phenomena occupy many researchers and are the topics of this section.

An Overview

According to Ransbotham (2016), financial advising is typically considered a knowledge-intensive task. As robot (robo) advisors provide personalized support for individuals, the costs of such services go down. This leads more people to demand such services, eventually freeing more humans to address advanced financial issues. Robo advisors may also cause some people to lose their jobs.

Some authors argue that the automation segment, which is related to cognitive computing and AI, will accelerate what is called *polarization* of the labor market in the future. This entails significant job growth in the top *and* bottom tiers of the labor market but losses in the middle. Jobs requiring low but specialized skills, such as personal care, are continuing to grow. Similarly, jobs that require very high skill, such as graphics design work, and so on, are also growing. But jobs that require "middle skills" such as specialized knowledge that is applied over and over with some adaptation, are at the greatest risk of disappearing. Sometimes technology disintermediates itself! For example, IBM Watson Analytics now includes querying capabilities to begin asking questions that an intelligent system professional previously asked and, obviously, providing answers. Other analytics-as-a service offerings with similar services may result in a need for fewer people to be proficient at using analytics software.

A report by *The Economist* notes that even if AI does not replace workers directly, it will certainly require employees to acquire new skills to keep their jobs. Market disruption is always uncomfortable. The next few years will provide excellent opportunities for intelligent technology professionals to shape the future.

Are Intelligent Systems Going to Take Jobs—My Job?

Tesla's Elon Musk envisions AI-based autonomous driving trucks all over the world within 10 years. There will be convoys of such trucks, each of which will follow a lead truck. Trucks will be electrical, economical, and pollution free. In addition, there will be fewer accidents—sounds great! But what about thousands of drivers who will lose their jobs? What about many thousands of employees in truck stops who will lose their jobs as well? The same scenario could happen in many other industries. Amazon has opened its first Go, a cashierless physical store. They plan 3,000 more in a few years. The post office in some countries already distributes mail using autonomous vehicles. In short, there is a chance for massive unemployment.

Example: Pilots at FedEx

FedEx has a fleet of close to 1,000 airplanes flying globally. According to Frank Tode, editor and publisher of *The Robot Report*, FedEx hopes that around 2020 the company will have one global pilot center with three or four pilots who will operate the entire FedEx fleet.

Foxcom, an iPhone manufacturer in Taiwan, had planned to replace almost all of its employees (60,000) in Taiwan with robots (Botton, 2016). The company already produced 10,000 robots for this purpose.

INTELLIGENT SYSTEMS MAY CREATE MASSIVE JOB LOSSES The debate regarding technology taking jobs has been going on since the beginning of the industrial revolution. The issue regarding intelligent systems is strongly debated now due to the following:

- They are moving very fast.
- They may take a large variety of jobs, including many white-collar and nonphysical jobs.
- Their comparative advantage over manual labor is very large and growing rapidly (see Figure 2.2 in Chapter 2).
- They are already taking some professional jobs from financial advisors, paralegals, and medical specialists.
- The capabilities of AI are growing rapidly.
- In Russia, robots are already teaching mathematics in schools (some do a better job than humans). Just think about what could happen to the teaching profession.

AI Puts Many Jobs at Risk

For the potential impact of AI on jobs, see Dormehl (2017), who explores the possibility of creative intelligent machines. For example, McKinsey's study estimates that AI is poised to take over 30 percent of all bank jobs in the near future. The study also predicts that robots will take 800 million jobs worldwide by 2030 (Information Management News, 2017).

To research the potential danger of job loss, McKinsey & Company divided jobs into 2,000 distinct work activities, such as greeting customers and answering questions about products, which retail salespeople do. Its researchers (see Chui et al., 2015) found that 45 percent of all 2,000 activities could be economically and physically automated. The activities include physical, cognitive, and social types.

While autonomous vehicles are not taking jobs, yet, they will take jobs from taxi drivers, Uber, and similar companies' drivers. Also, bus drivers may lose their jobs. Other jobs that have already been replaced by intelligent systems are listed in Application Case 14.2.

Application Case 14.2 White-Collar Jobs That Robots Have Already Taken

While it may be sometime before FedEx will have pilotless airplanes and schools will have no human teachers, some jobs, according to Sherman (2015), have already been taken by robots. They include:

- **Online marketers.** Using NLP, companies are automatically developing marketing ads and e-mails that influence people to buy (robo marketers). These are based on a dialog with potential buyers and on an automatic database search of historical cases. "Who needs an online marketer that may have inferior, biased, or incomplete knowledge?"
- **Financial analysts and advisors.** As was described in Chapter 12, robo advisors are all over the scene. Equipped with the ability to deal with Big Data in real time and conduct predictive analysis in seconds, these programs are liked by investors who pay about one-tenth of what human advisors charge. Furthermore, robo advisors can personalize recommendations.

- **Anesthesiologists, diagnosticians, and surgeons.** The medical field seems to be immune from AI. This is not the case. Expert systems for diagnosis have been in place for about 40 years. The FDA has already approved the J&J Sedasys system for delivery of low-level anesthesia in surgeries, such as colonoscopies. IBM's Watson has demonstrated a far more accurate diagnosis in lung disease cases than humans (90% vs. 50%). Finally, surgeons already use automated machines in some invasive procedures.
- **Financial and sports reporters.** These jobs involve gathering information, interviewing people, answering questions, analyzing the material, and writing reports. The Associated Press (AP) has experimented with AI machines since 2014. Results so far are virtually error and bias free (and no fake news!).

Palmer (2017) reported an additional five jobs in danger, including middle management, commodity salespeople, report writers, accountants and bookkeepers, and some types of doctors.

McFarland (2017b) lists as high-risk jobs cashiers, toll booth operators, fast-food employees, and drivers. Low-risk jobs include nurses, doctors, dentists, youth sport coaches, and social workers.

Questions for Case 14.2

1. Watch the 4:22 min. video about an interview with Palmer, at **linkedin.com/pulse/5-jobs-robots-take-first-shelly-palmer/**. Discuss some of the assertions made regarding doctors.
2. Discuss the possibility of your checkup by a robot-diagnostician. How would you feel?
3. With the bombardment of fake news and their biased creators, it may be wise to replace all of them by intelligent machines. Discuss such a possibility.
4. You are a defendant in a crime you did not commit. Would you prefer a traditional lawyer or one equipped with an AI e-discovery machine? Why?

Sources: Compiled from E. Sherman. (2015, February 25). "5 White-Collar Jobs Robots Already Have Taken." **Fortune.com**. **fortune.com/2015/02/25/5-jobs-that-robots-already-are-taking** (accessed April 2018); S. Palmer. (2017, February 26). "The 5 Jobs Robots Will Take First." *Shelly Palmer*.

Let us look at some other studies. A 2016 study done in the United Kingdom predicted that robots will take 50 percent of all jobs by 2026. Egan (2015) reports that robots already threaten the following jobs: marketers, toll booth operators and cashiers, customer service, financial brokers, journalists, lawyers, and phone workers. Note that automation may affect portions of almost all jobs to a greater or lesser degree. Experts estimate that about 80 percent of IT jobs may be eliminated by AI.

According to Manyika et al. (2017), automation is spreading because "robots are also increasingly capable of accomplishing activities that include cognitive capabilities once considered too difficult to automate successfully, such as making tacit judgments, sensing emotion, or even driving."

Given all this, you may wonder whether your job is at risk.

Which Jobs Are Most in Danger? Which Ones Are Safe?

If want to know about your job, it obviously depends on the type of job you are holding. Oxford University in the United Kingdom looked at 700 jobs and ranked them from zero (no risk of automation) to 1 (very high risk of automation). Straus (2014) provided a list of the top 100 most at-risk jobs (all above 0.95) and the 100 jobs with the lowest risk (with 0.02 or less). The top 10 "safe" and the 10 at risk are listed in Table 14.1.

A 2017 study conducted by the Bank of England found that almost half of the U.K. jobs (15 million out of 33.7 million) are at risk of loss within 20 years. Creative robots are the greatest threat because they can learn and increase their capabilities. While in the past, automation may not have decreased the total number of jobs, this time the situation may be different.

A side effect of this situation may be that workers will have less income while the owners of robots will have a larger income. (This is why Bill Gates suggested taxing the robots and their owners.)

TABLE 14.1 Ten Top Safe and at Risk Occupations

Probability of Job Loss
Low-Risk Jobs
0.0036 First-Line supervisors of firefighting and prevention workers
0.0036 Oral and maxillofacial surgeons
0.0035 Healthcare social workers
0.0035 Orthotists and prosthetists
0.0033 Audiologists
0.0031 Mental health and substance abuse social workers
0.0030 Emergency management directors
0.0030 First-Line supervisors of mechanics, installers, and repairers
0.0028 Recreational therapists
High-Risk Jobs
0.99 Telemarketers
0.99 Title examiners, abstractors, and searchers
0.99 Sewers, hand
0.99 Mathematical technicians
0.99 Insurance underwriters
0.99 Watch repairer
0.99 Cargo and freight agents
0.99 Tax preparers
0.99 Photographic process workers and processing machine operators
0.99 New account clerks

Source: Based on Straus (2014) Straus, R.R. "Will You Be Replaced by a Robot? We Reveal the 100 Occupations Judged Most and Least at Risk of Automation." **ThisisMoney.com**, May 31, 2014. **thisismoney.co.uk/money/news/article-2642880/Table-700-jobs-reveals-professions-likely-replaced-robots.html**

SOME MORE JOB LOSSES OBSERVATIONS

- Kelly (2018) predicts that robots could eliminate many Las Vegas jobs. And indeed, in many casinos worldwide, you can play several traditional games on machines.
- People with doctoral degrees have a 13 percent chance of being replaced by robots and AI versus 74 percent for those with only a high school education (Kelly, 2018).
- Women will lose more jobs to automation than men (Krauth, 2018).

Intelligent Systems May Actually Add Jobs

Despite the fear, uncertainty, and panic related to job losses, many reports contradict this. Here are some examples: de Vos (2018) reported that AI will create 2.3 million jobs in 2020 while eliminating 1.8 million. Also, one needs to consider the great benefits of AI and the fact that human and machine intelligence will complement each other in many jobs. Also, AI will increase international trade, adding more jobs. de Vos also cites studies that show the creation of jobs due to equipment maintenance and service that cannot be automated. The following are predictions on both sides of the issue:

- A PricewaterhouseCoopers (PwC) study forecast that robots will bolster U.K. economic growth. So, even though robots could destroy about 7 million jobs in the United Kingdom, they will create at least 7 million new jobs and probably more over 20 years (Burden, 2018).

- IBM's new deep learning service may help save IT jobs.
- There is a shortage of millions of skilled workers (e.g., about 50,000 truck drivers in the United States), so automation will reduce millions of unfilled positions.
- Korolov (2016) claims that there is plenty of work, especially for people who keep up with technology and broaden their skills.
- Gartner Inc. predicts that by 2020, AI will create more jobs than it eliminates Singh, (2017b).
- Wilson et al. (2017) report on new categories of human jobs that have been created by AI.
- Some believe that there will be a total of increase in jobs due to AI-induced innovations.
- It was estimated that in 2018 there would be over 490,000 jobs open for data scientists, but only 200,000 scientists will be available. However, in the long run, AI and machine learning may replace most data scientists (Perez, 2017).
- Violino (2018) contradicts those who claim that there is a huge fear among employees regarding job loss, saying that most workers see robots as an aid to their jobs. See also Leggatt (2017).

Note: When this book went to press, there was a shortage of IT employees (several million in the United States). Automation can alleviate this shortage. Note that a study reported by Weldon (2017b) showed that most workers actually welcome the impact on jobs by AI and automation. As a final note, Guha (2017) provides a view of work and AI as a vision of "despair, hope, and liberation." He concludes that AI can liberate work—it is a historical opportunity.

Jobs and the Nature of Work Will Change

While you may not lose your job, intelligent applications may change it. One aspect of this change is that low-skill jobs will be taken by machines, but high-skill jobs may not. Therefore, jobs may be redesigned either to be low skilled in order to be automated, or to be high skilled so that they will be executed exclusively by humans. In addition, there will be many jobs where people and machines will work together as a team.

Changes in jobs and business processes will impact training, innovation, wages, and the nature of work itself. Manyika (2017) and Manyika et al. (2017) of McKinsey & Company analyzed the shifts that can be fundamental, and arrived at the following conclusions:

- Many activities done by humans will have the potential to be automated.
- Productivity growth from robotics, AI, and machine learning will be tripled compared to pre-2015.
- AI will create many new jobs paying high salaries.
- Since more than half the world is still offline, the changes will not be too rapid.

Example: Skills of Data Scientists Will Change

According to Thusoo (2017) of the McKinsey Global Institutes study group, there will be a shortage of 250,000 data scientists by 2024. There will be a need to retrain or train scientists so they can deal with intelligent technologies and the changes in data science and in solving related real-world problems. Thus, proper education must evolve. The job requirements of data scientists are already changing. The scientists will need to know how to apply machine learning and intelligent technologies to build IoT and other useful systems. New algorithms improve operations and security, and data platforms are changing to fit new jobs.

Snyder (2017) found that 85 percent of executives know that intelligent technologies will impact their workforce within five years, and 79 percent expect the current

skill sets to be restructured. They also expect 79 percent productivity improvement. Employees fear that intelligent systems will take over some of their activities, but they hope that intelligent systems will also help with their work.

TIPS FOR SUCCESS A McKinsey study of 3,000 executives (Bughin, McCarthy, and Chui, 2017) reports the following success tips for implementing AI provided by the executives:

- Digital capabilities need to come before AI.
- Machine learning is powerful, but it is not the solution to all problems.
- Do not put technology teams solely in charge of intelligent technologies.
- Adding a business partner may help with AI-based projects.
- Prioritize a portfolio approach to AI initiatives.
- The biggest challenges will be people and business processes.
- Not every business is using intelligent systems, but almost all those that use them increase income and profit.
- Top leadership support is necessary for a transformation to AI.

DEALING WITH THE CHANGES IN JOBS AND THE NATURE OF WORK Manyika (2017) made the following suggestions for policymakers:

1. Use learning and education to facilitate the change.
2. Involve the private sector in enhancing training and retraining.
3. Have governments provide incentives to the private sector so employees can invest in improved human capital.
4. Encourage private and public sectors to create appropriate digital infrastructure.
5. Innovative income and wage schemes need to be developed.
6. Carefully plan the transition to the new work. Deal properly with displaced employees.
7. Properly handle new technology-enabled technologies.
8. Focus on new job creation, particularly digital jobs.
9. Properly capture the productivity increase opportunities.

Baird et al. (2017) of McKinsey & Company provide a video interview with industry experts discussing how to deal with the changing nature of work. Another exploration of the nature of work in the era of intelligent systems is provided by Crespo (2017). Chui et al. (2015) researched the impact of automation on redefining jobs and business processes, including the impact on wages, and the future of creativity. Finally, West (2018) provides a comprehensive study on the future of work as it is influenced by robotics and AI-driven automation.

Conclusion: Let's Be Optimistic!

Assuming that the disasters will not occur, then, as in the past, concerns about technology replacing many human jobs and reducing wages are hopefully exaggerated. Instead, intelligent technologies will clearly contribute to shorter work time for humans. Today, most people work long hours just for survival.

SECTION 14.6 REVIEW QUESTIONS

1. Summarize the arguments of why intelligent systems will take away many jobs.
2. Discuss why job losses may not be catastrophic.
3. How safe is your job? Be specific.

4. How may intelligent systems change jobs?
5. In what ways may work be changed?
6. Discuss some measures to deal with the changes brought by intelligent systems.
7. One of the areas of potential job loss is due to autonomous vehicles. Discuss the logic of this.

14.7 POTENTIAL DANGERS OF ROBOTS, AI, AND ANALYTICAL MODELING

During the period 2016–2018, we witnessed a heated debate regarding the future of AI and particularly robots. Dickson (2017) called the optimistic approach *Utopia* and the pessimistic one *Dystopia*. The debate began with the industrial revolution regarding automation, and it has accelerated because of the rapid technological innovations of AI. In Section 14.5, we presented one aspect of this debate, the impact on jobs. In the center of the debate is the prediction of when AI's capabilities to reason and make decisions will become similar or even superior to that of people. Furthermore, will such a development be beneficial or dangerous to society?

Position of AI Dystopia

The camp that supports this prediction includes well-known tech executives. Here are three of them:

- **Elon Musk:** "We need to be super careful with AI. Potentially more dangerous than nukes." (See the 10 min. video at **youtube.com/watch?v=SYqCbJ0AqR4**). Musk predicts that World War III will start because of AI. "Robots will kill us all, one day," he said in his several presentations.
- **Bill Gates:** "I am in the camp that is concerned about super intelligence. Musk and some others are on this and I don't understand why some people are not concerned." (Comments made on TV and interviews, several times). He also suggested taxing the manufacturers and users of robots and other AI machines.
- **Stephen Hawking:** The late scientist stated, "The development of full artificial intelligence could spell the end of the human race."

Many people are afraid of AI because they believe that computers will become smarter than we are. See Bostrom's video of his famous TED presentation at **youtube.com/watch?v=MnT1xgZgkpk**. See also Maguire (2017) for a discussion regarding learning robots and the risk of rebelling robots. For how robots can learn motor skills through trial and error, see the video at **youtube.com/watch?v=JeVppkoloXs/**. For more, see Pham (2018).

The AI Utopia's Position

A good place to begin for information on this position is to watch the 26 min. documentary video on the future of AI at **youtube.com/watch?v=UzT3Tkwx17A**. This video concentrates on the contribution of AI to the quality of life. One example is crime fighting in Santa Cruz, California, where AI was able to predict where and when crimes will occur. Following the predictions, the police department has been planning its work strategies. The result is a 20 percent reduction in crime.

A second example is the prediction of the probability that a certain song will be a hit. The prediction helps both artists and managers to plan their activities. Great success has been made. In the future, AI is predicted to compose top songs.

Finally, there is a story about dating. The capabilities of AI enabled a scientist to find a perfect match in a population of 30,000 potential candidates.

A basic argument of the Utopianists expressed in interviews, TV lectures, and more, is that AI will support humans and enable innovations. AI also will partner with humans. The Utopians believe that as AI expands, humans will become more productive and will have time to do more innovative tasks. At the same time, more tasks will be fully automated. Prices of products and services will drop and the quality of life will increase.

At one point, we may achieve a fully automated and self-sustaining economy. Ultimately, people will not have to work at all to make a living.

A leading proponent of AI benefits is Mark Zuckerberg of Facebook. He is in a heated debate with Elon Musk (CEO of Tesla Corp), the unofficial leader of the Dystopia camp of believers. Zuckerberg criticized those that believe that AI will cause "doomsday scenarios" (see the next section). Musk claimed that Zuckerberg has a "limited understanding" of AI, and Zuckerberg answered by referring to his paper on AI that won an award at the "top computer vision conference." For details, see Vanian (2017).

SOME ISSUES RELATED TO THE UTOPIA Several issues are related to the Utopianists' position. Here are three examples:

1. AI will be so great that people will have a problem of what to do with their free time. If you have not yet seen Disney's Wall-E movie, go and see it. It shows how humans are served by robots. Dennis Hassabis, a strong proponent of Utopia (from Deep Mind, an AI company), believes that AI will one day help people have a better life by understanding what makes humans unique, what the mysteries of the mind are, and how to enjoy creativity.
2. The road to AI Utopia could be rocky, for example, there will be impacts on jobs and work. It will take time to stabilize and adjust work and life of living with robots, chatbots, and other AI applications.
3. One day we will not drive anymore and there may not be human financial advisors; everything will be different, and the changes may be rapid and turbulent and we may even face disasters, as projected by the Dystopia camp.

The Open AI Project and the Friendly AI

To prepare against the unintended action of robotics and AI, Elon Musk and others have created Open AI, a nonprofit organization. With the unintended potential danger in mind, Musk and others created a nonprofit AI research company endowed with $1 billion. The major objective is to enact the path to safe artificial general intelligence (AGI). As you recall from Chapter 1, AGI is not here yet, but it is coming.

The plan of Open AI is to build safe AGI and ensure that its benefits will be evenly distributed. The research results are published in top journals. In addition, Open AI creates open source software tools. The organization has a blog and it disseminates important AI news. For details, see **openai.com**.

THE FRIENDLY AI Eliezer Yudkowsky, a cofounder of the Machine Intelligence Research Institute, developed the idea of *friendly AI*, according to which AI machines should be designed so that they will benefit humans rather than harm them (i.e., use a system of checks and balances in designing the AI capabilities). For details, see Sherman (2018), and view a fascinating 1:29:55 min. video by Yudkowsky (2016) at **youtube.com/watch?v=EUjc1WuyPT8**.

CONCLUSION It is difficult to know what will happen in the future. But some actions are already being taken to prevent a disaster. For example, several major companies have declared that they will not produce or support killer robots.

The O'Neil Claim of Potential Analytics' Dangers

Managers and data science professionals should be aware of the social and long-term effects of mathematical models and algorithms. Cathy O'Neil, a Harvard PhD in mathematics who worked in finance and the data science industry, expressed her experiences and observations in the popular book *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. We suggest you read the book, or at least visit the author's blog site at **mathbabe.org/**. The blog site highlights social issues related to analytics. A good summary/review of the book is available at **knowledge.wharton.upenn.edu/article/rogue-algorithms-dark-side-big-data/**.

In her book, O'Neil (2016) argues that models must satisfy three conditions. First, they must be transparent. That is, if the model is not understandable, its application can lead to unintended consequences.

Second, the model must have clear quantifiable objectives. For example, the celebrated application of analytics in the book and movie *Moneyball* includes a model that was aimed at increasing the number of financial wins. And the proposed input measures were well understandable. Rather than using the more commonly reported measure "run base in" (RBI), the analyst in *Moneyball* proposed and used on-base percentage and other measures (which were also easily calculated and understood by anyone with basic math skills). On the other hand, models built to assess the risk of mortgage-backed securities when no one fully understood the underlying assumptions of collateralized securities, but financial traders were trading, have been blamed for leading the financial crisis of 2008.

The third requirement is that the models must have a self-correcting mechanism and a process in place so that they are audited regularly and new inputs and outputs are constantly being considered. This third issue is particularly critical in applying models in social settings. Otherwise, the models perpetuate the faulty assumptions inherent in the initial modeling stage. O'Neil discusses several situations where such is the case. For example, she describes the models built in the United States to identify underperforming teachers and reward better teachers. Some of these models utilized the test scores of the pupils to assess the teachers. O'Neil cited several examples where the models were used to fire "underperforming" teachers even though those teachers were loved by the students and parents. Similarly, models are used to optimize the scheduling of workers in many organizations. These schedules may have been developed to meet seasonal and daily demand variations, but the models do not take into account the deleterious impacts of such variability in schedules on the families of these usually lower-income workers. Other such examples include credit score assessment models that are based on historical profiles and thus may negatively impact minorities. Without mechanisms to audit such models and their unintended effects, they can do more harm than good in the long term. So, model builders need to consider such concerns.

Note: In May 2018, General Data Protection Regulation (GDPR) became effective in the European Union. It includes the need to explain data. According to Civin (2018), an explainable AI could reduce the impact of biased algorithms.

A comment: There is evidence that in some cases O'Neil's claims are valid, and therefore model builders and implementers must pay attention to the issues. However, in general, analytics are properly designed and bring considerable benefits to society. Furthermore, analytical models increase the competitiveness of companies and countries, creating many highly paid jobs. In many cases, companies have social responsibility policies that minimize biases and inequality. Finally, as Weldon (2017a) observed, algorithms and AI can be seen as great equalizers in bringing services that were traditionally reserved for a privileged few, to everyone.

SECTION 14.7 REVIEW QUESTIONS

1. Summarize the major arguments of the Utopia camp.
2. Summarize the major arguments of the Dystopia camp.
3. What is the friendly AI?
4. What is Open AI? Relate it to the dystopia vision.
5. What are the potential risks in using modeling and analytics?

14.8 RELEVANT TECHNOLOGY TRENDS

As we near the last section of this book that discusses some aspects of the future of intelligent systems, it is worthwhile to describe some of the technology trends that will shape this future. Unfortunately, there are hundreds of technology trends relevant to the content of this book. The reason is that there are hundreds of variations of analytics, Big Data tools, AI, machine learning, IoT robotics, and other intelligent systems. Therefore, we provide here only a sample of technology trends. We divide this section into the following subsections:

- Gartner's 2018 and 2019 lists.
- List of technology trends in intelligent systems.
- Ambient computing.

Gartner's Top Strategic Technology Trends for 2018 and 2019

Gartner Inc. is a top technology research organization and consultant as well as an organizer of an annual technology symposium attended by over 23,000 people (Gartner Symposium IT expo). It provides an annual prediction of the technologies that it thinks will impact most organizations. The 2018 and 2019 lists of trends includes 10 items each, most of which relate directly to the content of our book.

The summary of the 2018 list is shown in Figure 14.3. It was extracted from Gartner's press release of October 4, 2017, which is available at **gartner.com/newsroom/id/3812063**. The essentials are provided in a video (5:36 min.) at **youtube.com/watch?v=TPbKyD2bAR4**.

GARTNER'S 2018 AND 2019 LISTS The following is extracted from **gartner.com/newsroom/id/3812063**, for 2018, and from Weldon (2018), for 2019.

1. **AI Foundation and Development.** Advanced AI systems that support decision making, some of which are autonomous, and other AI systems are developed in conjunction with analytics and data science.
2. **Intelligent Apps and Analytics.** Almost all IT systems will include AI in the next few years. See[**gartner.com/smarterwithgartner/the-cios-journey-to-artificial-intelligence/**].
3. **Intelligent and Autonomous Things.** Utilizing the IoT capabilities, there will be an explosion of autonomous vehicles and a significant increase of other intelligent things (e.g., smart homes and factories where robots are assembling robots).
4. **Digital Twin.** A digital twin, see [**gartner.com/smarterwithgartner/prepare-for-the-impact-of-digital-twins/**], refers to digital representations of real-world objects and systems. This includes mainly IoT systems with 20 billion connected things in two to three years.
5. **Empowered Cloud (Cloud to the Edge).** In Edge computing, information collection, processing, and delivery are conducted closer to the sources of the information.

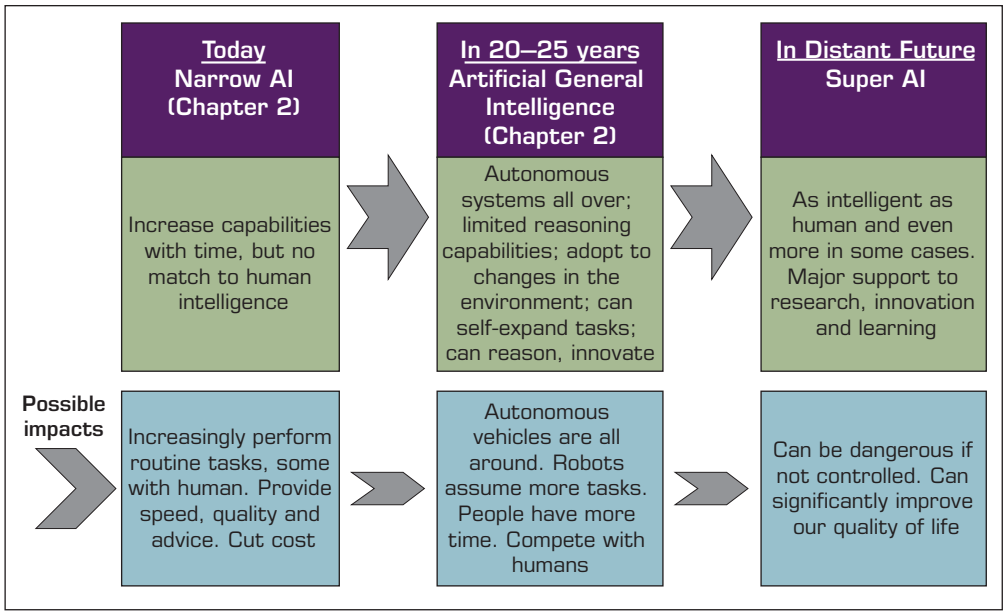


FIGURE 14.3 **Predict the future of AI (Drawn by E.Turban)**

6. **Conversational Human-Machine Platforms.** These platforms already facilitate natural language interactions, resulting in improved collaboration. These include smart collaborative spaces..
7. **Immersive Experience.** These systems change the manner in which people can see and perceive the world (e.g., augmented reality). See [**gartner.com/smarterwithgartner/transform-business-outcomes-with-immersive-technology/**].
8. **Blockchain.** Blockchain technologies [**gartner.com/smarterwithgartner/are-you-ready-for-blockchain-infographic/**] offer a radical platform for increased security and trust, significantly improving business transactions.
9. **Augmented Analytics.** Using machine learning enables this technology to focus on transformation of analytics, so it will be better shared and consumed. This will facilitate data preparation management and analysis to improve decision support.
10. **Others.** These include smart collaboration space, Quantum computing, digital and ethical privacy, and adopting risks and trust.

Other Predictions Regarding Technology Trends

- The IEEE computer society also has 10 top predictions for 2018. **computer.org/web/pressroom/top-technology-trends-2018**. The list includes deep learning, industrial IoT, robotics, assisted transportation, augmented (assisted) reality, blockchain, and digital currencies.
- Newman (2018) provides a list of 18 tech trends at CES 2018. These are related to displays at CES.
- The potential business application and value for several analytics and AI technologies based on studies of 400 real-world cases done at McKinsey & Company is available as interactive data visualization at **mckinsey.com/featured-insights/artificial-intelligence/visualizing-the-uses-and-potential-impact-of-ai-and-other-analytics/** (posted April 2018).

- Top 10 trends for analytics in 2018 are provided by Smith (2018). The list is fairly technical in nature. It includes "Data Gravity will accelerate to the cloud," "Insight-as-a-service will rise," and "End-to-end cloud analytics will emerge."
- Top 10 AI technology trends for 2018 as envisioned by Rao et al. (2017) include "Deep reinforcement learning: interacting with the environment to solve business problems" and "Explainable AI: understanding the black box."
- For seven data and analytical trends, see **datameer.com/blog/seven-data-analytics-trends-2018/**.
- Computers will learn to think and think to learn.
- Robots will replace humans in more nonphysical and cognitive roles.
- Intelligent augmentation is part of the narrow AI (Chapter 1) and will continue to control new AI applications.
- Edge computing was cited by Gartner, but it has much more value that may not be related to the "cloud." The technology will have a major impact on the future of data centers. For details, see Sykes (2018a). Note that most of the new capabilities for the "cloud" exist in the use of the "Edge." For further information, visit Wikipedia. Edge AI enhancements will excel in supporting machine learning and augmented reality.

Sommer (2017) lists the following:

- Data literacy will spread both in organizations and in society.
- Information points will be connected via hybrid multi-cloud systems.
- The mystery of rural networks will be exposed by deep learning theory.
- Self-service systems will use data catalogs as their frontier.
- Need to focus on Application Programming Interfaces (APIs).
- Analytics become conversational (e.g., via chatbots).
- Analytics will include immersive capabilities.
- Using augmented intelligence users will be turned to participants.
- For 11 top trends that drove business intelligence in 2018, see Sommer (2017).
- For six data analytics trends in 2018, see Olavsrud (2018).
- For robotics trends in 2018, see Chapman (2018).
- For 10 predictions of intelligent systems, see Press (2017).

Summary: Impact on AI and Analytics

Now that you have seen the many technologies trends for the future, you may also want to see when they will impact AI. Figure 14.3 illustrates the long-term projection of AI. The future is divided into three sections: today, in about 20 years, and in a distant future.

The future of BI and analytics is illustrated in Figure 14.4. Some additional predictions are intelligent analytics, insight-as-a-service, and data cataloging. Finally, we describe one technology in more detail. It may impact both analytics and AI.

Ambient Computing (Intelligence)

Closely related to the IoT, chatbots, smart homes, analytics, sensors and "things" are included in the concept of **ambient computing** (or paradigm computing). It has several definitions, but essentially it refers to electronic environments (e.g., network devices such as sensors) that are sensitive and responsive to people and their environments. So ambient devices can support people in whatever task they are doing. Once sensing their surroundings, the devices provide different input/output methods that depend on the configuration of situations (e.g., what people are doing at a given

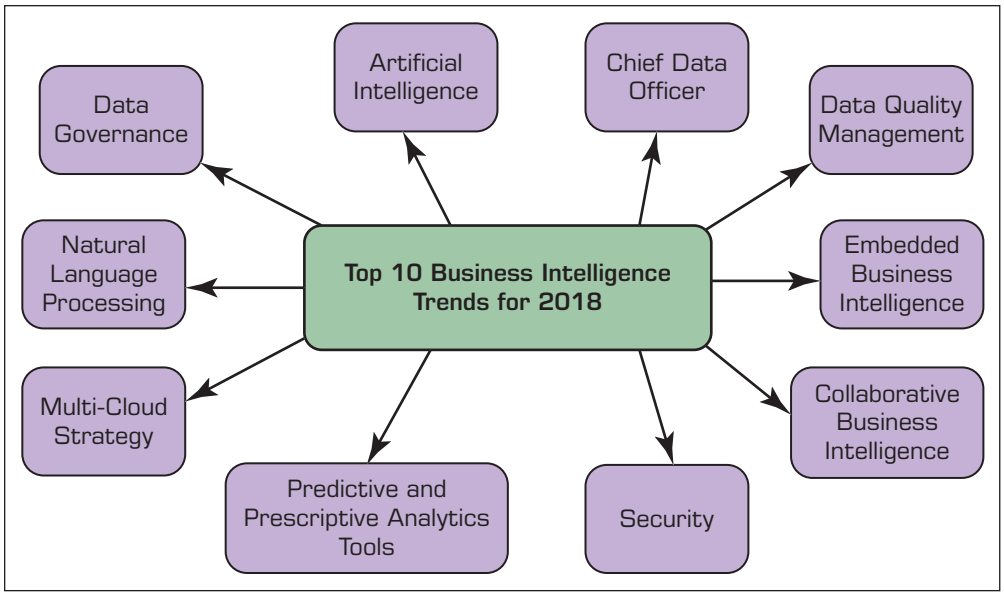


FIGURE 14.4 **Future of Analytics** *Source:* "Analytics and BI Trends", Datapine, in Top 10 Analytics and Business Intelligence Trends for 2018, Business Intelligence, Dec 13th 2017, © 2017, Used with permission.

time). In summary, everything in our life will be computerized and intelligent. The concept is based on previous research in the areas of pervasive computing, human–machine interaction, context awareness, profiling, personalization, and interaction design. For details, see **en.wikipedia.org/wiki/Ambient_intelligence** and Charara's (2018) guide.

POTENTIAL BENEFITS OF AMBIENT INTELLIGENCE While the concept is mostly futuristic, its characteristics and benefits are already envisioned. The networked devices can:

- Recognize individuals and other "things" and their context at any given time and place.
- Integrate into the environment and existing systems.
- Anticipate people's desires and needs without asking (e.g., context awareness).
- Deliver targeted services based on people's needs.
- Be flexible (i.e., can change their actions in response to people's needs or activities).
- Be invisible.

Many of the devices and services described in this book already exhibit some of the capabilities of ambient computing. Amazon's Alexa is probably currently the closest to the ambient concept. For details, see Kovach (2018). For more on ambient computing and its relationship to IoT and smart cities, see Konomi and Roussos (2016).

SECTION 14.8 REVIEW QUESTIONS

1. Identify three of the Gartner 10 that are mostly related to analytics and data science.
2. Identify three of the Gartner 10 that are most related to AI and machine learning.
3. Identify three of the Gartner 10 that are most related to IoT, sensors, and connectivity.

4. Identify three technologies related to analytics from the other predictions list and explore them in more detail. Write a report.
5. Identify three data science–related technologies from the long list and explore them in more detail. Write a report.
6. Identify three AI-related technologies from the long list and explore them in more detail. Write a report.
7. Describe ambient computing and its potential contribution to intelligent systems.

14.9 FUTURE OF INTELLIGENT SYSTEMS

There is a general agreement among AI experts that AI is going to change everything in our world for the better (e.g., see Lev-Ram [2017] and Violino [2017]). However, there are disagreements on when such changes will occur and what their impact is going to be. AI research is accelerating due to improvements in different related computer technologies (e.g., chips, IoT), improvements in intelligent methodologies and tools, the increased activities in high-tech companies that are striving to gain leadership in certain intelligent systems areas and firms that are investing billions of dollars in AI, the development of AI tools and methodologies, and much more. In this section, we first provide a presentation of what some major corporations are doing in the intelligent technologies field.

What Are the Major U.S. High-Tech Companies Doing in the Intelligent Technologies Field?

One way to predict the future of AI is to look at what the major companies are currently doing.

GOOGLE (ALPHABET) Google uses NLP in its Google Translate as well as in its search processes. It uses neural networks in its immersed databases (for pattern recognition) and for making decisions on them. In addition, Google uses other machine-learning algorithms for personalization advertising decisions. Google Assistant and Home are two applied projects that attracted considerable attention in CES 2018. Google Assistant is trying to dethrone Alexa. In addition, Google is most active in the autonomous vehicles field. Google purchased several AI companies and is conducting extensive research in the field. Google has a special team that attempts to provide Google AI speech dialog with a personality (see Eadicicco, 2017). Google DeepMind's AlphaGo is the machine that beat the game Go champions. Google is using machine learning for managing its huge databases and search strategies. Finally, Google is teaching its AI machines how people behave (e.g., cook, huy) by showing them film clips (see Gershgorn, 2017).

APPLE Apple is known to secretly be working on several AI projects. The most known is its Siri chatbot, which is embedded in several of its products (e.g., iPhone). In 2016, Apple acquired a machine-learning company, Turi. While lagging behind Google, Amazon, and Microsoft, Apple is rapidly closing the gap, using acquisitions and extensive research and development. Apple acquired companies in speech recognition (Vocal!), image recognition (Perception), and facial expression recognition (Emotion). Thus, Apple is becoming a leader in AI. With several hundred millions of Siri users and new acquisitions in AI, Apple is charging forward rapidly.

FACEBOOK Mark Zuckerberg, Facebook's CEO, is a major believer in the future of AI. In addition to his personal investments in AI, he hired Yann LeCun, a deep-learning pioneer,

to lead AI research in the company. LeCun created a special Facebook unit that identifies important AI developments and incorporates them into Facebook's products. Facebook invested billions of dollars in AI. With Facebook, AI goes mainstream. With its over 2 billion users, Facebook is spreading its AI applications globally.

MICROSOFT Microsoft is very active in all AI technology research. In 2017, it acquired Maluuba, a start-up that specializes in deep learning and NLP. Some believe that this acquisition will help Microsoft outperform both Facebook and Google in the areas of speech and image recognition. Maluuba excels in reading and comprehending text with near human capabilities in its virtual personal assistant, Cortana. This assistant helps people deal with e-mail and messaging difficulties. The AI will examine the content of messages and any stored documents and advice for what actions to take. For a comprehensive video about AI today and tomorrow by Stanford University, watch a 74 min. seminar at **youtube.com/watch?v=wJqf17bZQsY**.

IBM IBM entered robotics as early as 1973. By 1980, it had developed the QS-1; by 1977, it had developed Deep Blue; and by 2014, a mature IBM Watson entered the scene. IBM is also known for its artificial brain project. (For Blue Brain, see **artificialbrains.com/blue-brain-project**.) IBM is also known for its Deep QA project.

IBM is very active in AI research, especially in the area of cognitive computing; see Chapter 6 and **research.ibm.com/ai/**. IBM Watson was developed in collaboration with MIT AI labs.

Some other current projects focus on distributed deep learning software, creation of music and movie trailers by machines, gesture recognition, combining AI and IoT (e.g., embodied cognition), and medical applications supported by Watson (cognitive care, e.g., cancer detection, mental health care, and visually impaired people). IBM Watson is already considered the strongest applied brand of AI. One billion users were expected to use it in 2018, gaining substantial benefits from its applications.

AI Research Activities in China

AI research is done in many countries, notably Germany, Japan, France, the United Kingdom, and India. But most research outside the United States is done in China. China plans to be the world leader in AI, and its government is strongly supporting the activities of many AI companies. As you may recall from Chapter 1, Vladimir Putin has said that whoever leads AI will control the world. And, indeed, China plans to be that leader by 2030. The country plans an AI industry of $150 billion.

Among the many companies that are engaged in AI, three are investing billions of dollars, employing thousands of AI experts and robotic engineers, and acquiring global talents in AI. The three companies are *Alibaba Group, Tencent*, and *Baidu*. AI is already the priority of the Chinese government. In a cover story in *Fortune*, Lashinsky (2018) describes and analyzes the competition between Tencent and Alibaba.

TENCENT This giant e-commerce company has created a huge AI lab to manage its AI activities. The goal is to improve AI capabilities and support decision making in the following areas: computer vision, NLP, speech recognition, machine learning, and chatbots. AI is already embedded in over 100 Tencent products, including WeChat and QQ. A well-known AI slogan in China is "Juey, GO AI." Tencent supports the robotic company UBTech Alpha. Tencent is the world's largest Internet company, and AI improves its operations. Another slogan is "AI in all." The company has a lab in Bellevue, Washington. Healthcare is a main research priority there. For more on AI at Tencent, see Marr (2018).

BAIDU Baidu started NLP research five years before Google to improve its search engine capabilities. The company is located in the Silicon Valley, Seattle, and Beijing. Baidu has several products. One is Duer OS, a voice assistant that is embedded in more than 100 brands of appliances in several countries. The product is now optimized for smartphones. Baidu is also working on autonomous vehicles. Finally, the company promotes facial recognition in the enterprise (replacing ID badges). Baidu's AI is growing but still much smaller than that of Alibaba.

ALIBABA The world's largest e-commerce company and the provider of cloud computing and IoT platforms, Alibaba is active in AI projects and is an investor in AI companies, such as in the face recognition giant SenseTime. Alibaba has developed a methodology for conducting AI, which is described in Application Case 14.3.

Application Case 14.3 How Alibaba.com Is Conducting AI

Alibaba has developed a cloud-based model known as ET Brain **alibabacloud.com/et**. The logic is that today and in the near future, we are and will be doing business in the cloud computing environment. Content, knowledge, and data are in the cloud, and Alibaba is both a user and a provider of iCloud. The ET Brain model is illustrated in Figure 14.6.

ETBrain consists of three parts: technologies, capabilities, and applications. Technologies include Big Data and analytic processing, neural networks, video recognition analysis, and machine learning. These technologies provide four major capabilities such as cognitive perception, reasoning, real-time decision making, and machine learning (see the middle level in the figure). The capabilities drive a large amount of applications, such as e-commerce activities (both business-to-business and business-to-consumers), medical and health care, smart cities, agriculture, travel, finance, and aviation. All-in-all, it is a super-intelligent AI platform. The ET Brain is illustrated in a 26:29 min. video at **youtube.com/watch?v=QmkPDtQTarY**.

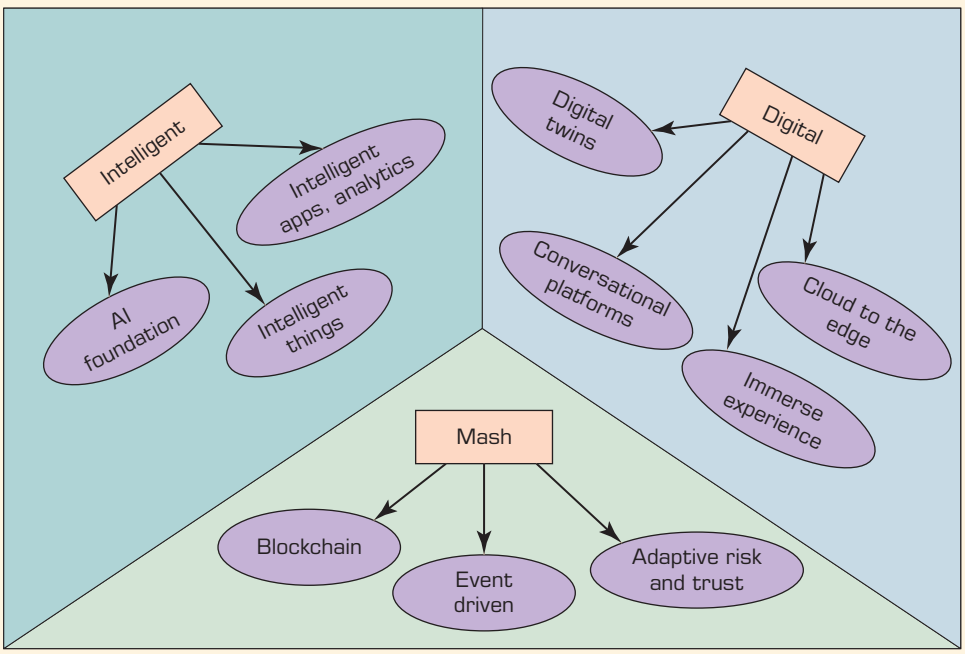


FIGURE 14.5 Gartner Prediction. Drawn by E. Turban

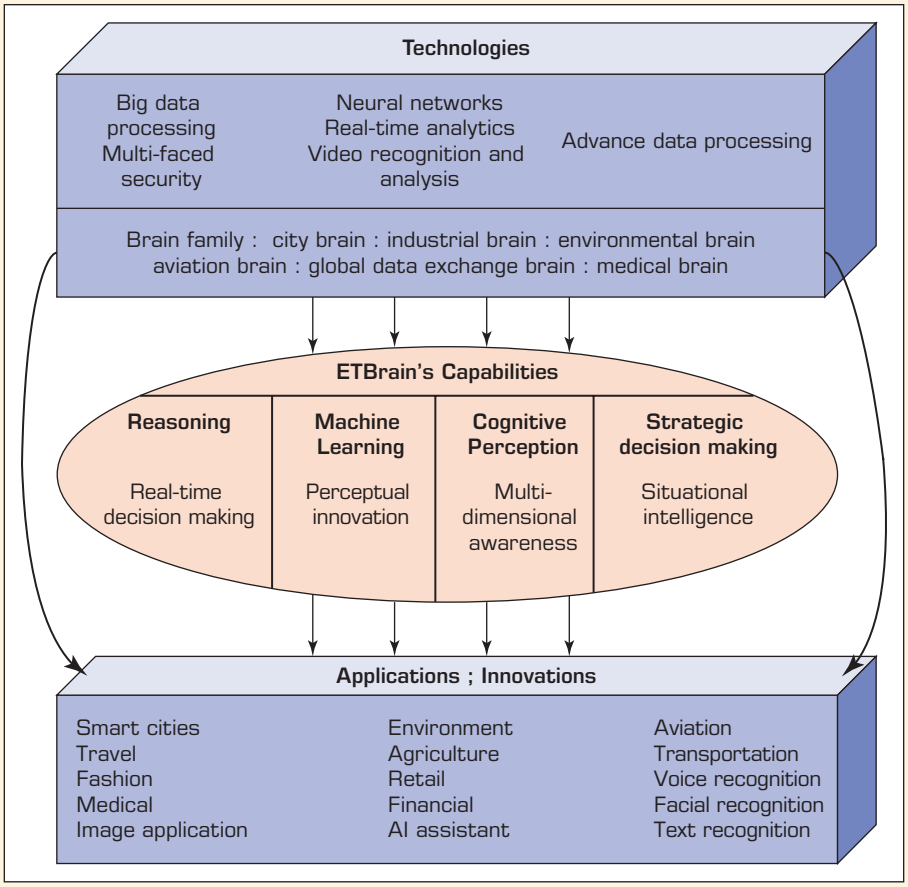


FIGURE 14.6 **Alibaba's ET Brain Model.** Drawn by E. Turban. Based on text at **Alibabacloud.com/et**

Alibaba's mission is to reach 2 billion consumers and to help 10 million businesses worldwide. To attain this mission, the company invested in seven research labs that focus on AI, machine learning, NLP, face (image) recognition, and network security. Alibaba is using AI to optimize its supply chain, personalize recommendations, and provide virtual personal assistants. Alibaba concentrates on several industries and on AI-supported bricks-and-mortar shopping. For example, in its AI office in Hong Kong, the company opened "Fashion AI," working with Guess Inc., helping shoppers to create an online ensemble while they are in a physical store. See **engadget.com/2018/07/04/guess-alibaba-ai-fashion-store/**. The company plans to rewire the world with AI (see Knight, 2018) and may control the world commerce.

Questions for Case 14.3

1. Relate cloud computing to AI at Alibaba.
2. Explain the logic of the ET Brain model.
3. Search the Web to find recent Alibaba activities in the AI field.
4. Read Lashinsky (2018). Why is Alibaba in such strong competition with Tencent?

Sources: Compiled from W. Knight. (2018, March 7). "Inside the Chinese Lab That Plans to Rewire the World with AI." *MIT Technology Review*; Marr, B. (2018, June 4). "Artificial Intelligence (AI) in China: The Amazing Ways Tencent Is Driving Its Adoption." *Forbes*; A. Lashinsky. (2018, June 21). "Alibaba v. Tencent: The Battle for Supremacy in China." *Fortune*. **alibabacloud.com/et**.

The U.S.–China Competition: Who Will Control AI?

At the moment, U.S. companies are ahead of Chinese companies. However, this situation may be changed in the future due to the huge investments in AI in China and the support provided by the Chinese government. Note that a major topic in the U.S.–China trade negotiations in 2018 centered on the use of technology by Chinese companies that employ U.S. knowledge and trade secrets.

The Largest Opportunity in Business

According to McCracken (2017), intelligent technologies provide the largest opportunity for tech companies since mobile computing. This is why tech giants and start-ups are trying to exploit AI. Desjardins (2017) provides an infographic about the future impact of AI that includes $15.7 trillion by 2030 in the form of productivity gains and increased consumer spending. By 2018, tech giants and others will invest $30 billion in research and development and $13.2 billion in start-ups. The largest improvement is expected in image and speech recognition products.

Note that despite their rivalry, Facebook, Amazon, Google, IBM, and Microsoft launched a partnership to research advancements and best practices in AI.

Conclusion

Now that you have completed reading this book you may ask, "What will happen to intelligent technologies in the future?" There will be a significant impact on business and quality of life. There will be changes, and they will be significant. With billions of dollars invested, mostly in AI, there will be advancements. Machines are getting smarter and smarter. For example, Alibaba's copywriting machine, which is based on deep learning and NLP, can generate 20,000 lines of text in one second. The machine is so smart that it passed the Turing test (Chapter 2), which means that it is smart like a human but can work much faster. We will now look at two areas: business and quality of life.

IMPACT ON BUSINESS According to Kurzer (2017), there might be challenges, but AI was expected to flourish as of 2018. There is very little doubt that we will see increased commercialization of AI, especially in marketing, financial services, manufacturing, and IT support. For example, the quality and nature of the customer experience could be improved, augmented by AI applications, and IoT. Kurzer also predicted that there will be more proactive processes rather than reactive ones. There will be more people-machine collaboration and while many jobs will be automated, many new ones will be created. There is going to be more conversational AI due to the increased capabilities of chatbots and personal assistants such as Alexa, Siri, and Google Assistant. Gartner predicted that by the end of this decade, people will have more conversations with machines than with their immediate family members (**gartner.com/smarterwithgartner/gartner-predicts-a-virtual-world-of-exponential-change/**). Another area with promising applications is *image recognition*. Google is a major force in both conversational and image recognition AI.

IMPACT ON QUALITY OF LIFE There will be impacts on life that will change the way we drive, eat, entertain, get services, learn, and fight.

A major area where AI intelligent systems have already made a stride is the healthcare field. Bernard Tyson, CEO of Kaiser Permanente, made the following public statement: "I don't think any physician should be practicing without AI assisting in their

practice. It's just impossible (otherwise) to pick up on patterns, to pick up on trends, to really monitor care." Editors (2018) report that smart solutions can improve quality of life indicators by 10 to 30 percent. (The longer we wait, the higher the percentage will be.) Among the indicators that they cite are: having longer and healthier lives, reducing greenhouse gas emissions, saving 200,000 lives worldwide over 10 years (thanks to self-driving cars), reducing the commute time for people (fewer traffic problems), increasing the number of jobs (e.g., by new technologies and more productive business environments), and providing better and more affordable housing.

Autonomous vehicles, including drones, will clearly change our lives for the better, and robots will be able to serve us (especially people who are elderly and those that are sick), entertain us, and if properly managed, be our companions. For an impact of AI in the future on society, watch the video at **youtube.com/watch?v=KZz6f-nCCN8/**.

What will the unintended results be? What if robots will kill us all? Well, that probably will never happen. People are smart enough to make sure that only good results will come from intelligent systems.

SECTION 14.9 REVIEW QUESTIONS

1. Describe the AI activities of major U.S. tech companies.
2. Describe the work by Chinese giant companies.
3. Describe Alibaba's approach to AI (The ET Brain model).

Chapter Highlights

- Intelligent systems can affect organizations in many ways as stand-alone systems, or integrated among themselves or with other computer-based information systems.
- The impact of analytics on individuals varies—it can be positive, neutral, or negative.
- Serious legal issues may develop with the introduction of intelligent systems; liability and privacy are the dominant problem areas.
- Many positive social implications can be expected from intelligent systems. These range from providing opportunities to people to lead the fight against terrorism. Quality of life, both at work and at home, is likely to improve as a result of the use of these technologies. Of course, there are potentially negative issues to be concerned about.
- Growth of intelligent systems is going to lead to major changes in industry structure and future employment.
- A major battle is brewing about who owns the user data that are being generated from the use of smartphones, cars, and so on.
- In deploying intelligent systems, it is necessary to consider legal, privacy, and ethical issues.
- Placing robots as coworkers in the work force raises legal and ethical issues.
- Intelligent technologies may impact business processes, organizational structure, and management practices.
- It may be necessary to create independent organizational units that deploy and manage intelligent systems.
- Intelligent systems may provide a considerable competitive advantage to their users.
- Intelligent systems may create massive unemployment mainly in routine and mid-management jobs.
- Eventually, intelligent system may cause unemployment even in skilled jobs. So retraining may be needed.
- Intelligent systems may result in restructuring many jobs notably through human-machine collaboration.
- Intelligent systems will create many new jobs that require specialized training.
- The use of intelligent systems automation may result in a shorter work week and a need to compensate those people who will lose their jobs.
- Some people are afraid of unintended consequences of having AI and robots. Machines will learn and may harm humans.

Key Terms

ambient computing computer ethics privacy

Questions for Discussion

1. Some say that analytics in general dehumanize managerial activities, and others say they do not. Discuss arguments for both points of view.
2. Diagnosing infections and prescribing pharmaceuticals are the weak points of many practicing physicians. It seems, therefore, that society would be better served if analytics-based diagnostic systems were used by more physicians. Answer the following questions:
   a. Why do you think such systems are used minimally by physicians?
   b. Assume that you are a hospital administrator whose physicians are salaried and report to you. What would you do to persuade them to use an intelligent system?
   c. If the potential benefits to society are so great, can society do something that will increase doctors' use of such intelligent systems?
3. What are some of the major privacy concerns in employing intelligent systems on mobile data?
4. Identify some cases of violations of user privacy from current literature and their impact on data science as a profession.
5. Some fear that robots and AI will kill all of us. Others disagree. Debate the issue.
6. Some claim that AI is overhyped. Debate the issue. Place a question on Quora and analyze five responses.
7. Some claim that AI may become a human rights issue (search for Safiya Noble). Discuss and debate.
8. Discuss the potential impact of the GDPR on privacy, security, and discrimination.
9. Discuss ethics and fairness in machine learning. Start by reading Pakzad (2018).
10. Should robots be taxed like workers? Read Morris (2017) and write about the pros and cons of the issue.

Exercises

1. Identify ethical issues related to managerial decision making. Search the Internet, join discussion groups/blogs, and read articles from the Internet. Prepare a report on your findings.
2. Search the Internet to find examples of how intelligent systems can facilitate activities such as empowerment, mass customization, and teamwork.
3. Investigate the American Bar Association's Technology Resource Center (**americanbar.org/groups/departments_offices/legal_technology_resources.html**) and **nolo.com**. What are the major legal and societal concerns regarding intelligent systems? How are they being dealt with?
4. Explore several sites related to healthcare (e.g., **WebMD.com**, who.int). Find issues related to AI and privacy. Write a report on how these sites suggest improving privacy.
5. Go to **Humanyze.com**. Review various case studies and summarize one interesting application of sensors in understanding social exchanges in organizations.
6. Research the issue of voice assistants and privacy protection. Start by reading Collins (2017) and Huff (2017).
7. Is granting advanced robots rights a good or bad idea? Read Kottasova (2018) for a start.
8. Face and voice recognition applications are mushrooming. Research the state of their regulation in a country of your choice. Use the United States if your country is not regulating.
9. Research the ethical issues of self-driving cars. Start by reading Himmelreich (2018).
10. Is your organization ready for AI? Research this issue and find all major activities that it includes.
11. Research the role of IoT as a tool for providing connectivity between sensors and analytics. Write a report.
12. Some people say that robots and chatbots may increase insurance risk and fees. Research this and write a report.
13. Watch the video at **youtube.com/watch?v=wwuovuCfDU/** and comment about the robot's potential impacts.
14. Research the issue stated in quotation marks: "When will robots rebel?" and "Will AI take control of the plant?" Start by reading Maguire (2017) and read **advancedmp.com/artificial-intelligence/**. Write a report.
15. Read Chui et al. (2016) and research the areas in which machines can replace humans and where they cannot (yet). Find changes since 2016. Write a report.
16. Watch the 3:38 min. video at **youtube.com/watch?v=781MlkxyqI/**. Relate it to Musk's predictions about robots reigning in this world (Section 14.7).
17. Read the SAS report on AI ethics at **sas.com/en_us/insights/articles/analytics/artificial-intelligence-ethics.html**. Comment on each of the three proposed steps. Also comment on the human-machine collaboration in problem solving.

References

Ainsworth, M. B. (2017, October). "Artificial Intelligence for Executives." *SAS White Paper, ai20for20executives.pdf*, October 2018.

Andronic, S. (2017, September 18). "5 Ways to Use Artificial Intelligence as a Competitive Advantage." **Moonoia.com**.

Anon. (2017, February 20). "Big Data and Data Sharing: Ethical Issues." *UK Data Service*. **ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf** (accessed July 2018).

Autor, D. H. (2016, August 15). "The Shifts—Great and Small—in Workplace Automation." *MIT Sloan Review*. **sloanreview.mit.edu/article/the-shifts-great-and-small-in-workplace-automation/** (accessed July 2018).

Baird, Z. et al. (2017, August). "The Evolution of Employment and Skills in the Age of AI." McKinsey Global Institute.

Baroudy, K., et al. (2018, March). "Unlocking Value from IoT Connectivity: Six Considerations for Choosing a Provider." *McKinsey & Company*.

Batra, G., A. Queirolo, & N. Santhanam. (2018, January). "Artificial Intelligence: The Time to Act Is Now." *McKinsey & Company*.

Bloomberg News. (2017, November 29). "Ethical Worries Are Marring Alphabet's AI Healthcare Initiative." *Information Management*.

Bossmann, J. (2016). "Top 9 Ethical Issues in Artificial Intelligence." *World Economic Forum*.

Botton, J. (2016, May 28). "Apple Supplier Foxconn Replaces 60,000 Humans with Robots in China." *Market Watch*.

Brynjolfsson, E., & A. McAfee. (2016). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. Boston, MA: W.W. Norton.

Bughin, J., B. McCarthy, & M. Chui. (2017, August 28). "A Survey of 3,000 Executives Reveals How Businesses Succeed with AI." *Harvard Business Review*.

Burden, E. (2018, July 16). "Robots Will Bolster U.K. Growth and Create New Jobs, PwC says." *Bloomberg News*.

Catliff, C. (2017, August 15). "Three Ways Your Business Can Leverage Artificial Intelligence." *The Globe and Mail*.

Chapman, S. (2018, January 16). "The Robotics Trends of 2018, According to Tharsus." *Global Manufacturing*.

Charara, S. (2018, January 4). "A Quick and Dirty Guide to Ambient Computing (and Who Is Winning So Far)." **Theambient.com**.

Chui, M., K. George, & M. Miremadi. (2017, July). "ACEO Action Plan for Workplace Automation." *McKinsey Quarterly*.

Chui, M., J. Manyika, & M. Miremadi. (2015, November). "Four Fundamentals of Workplace Automation." *McKinsey Quarterly*.

Chui, M., J. Manyika, & M. Miremadi. (2016, July). "Where Machines Could Replace Humans—and Where They Can't (Yet)." *McKinsey Quarterly*.

Civin, D. (2018, May 21). "Explainable AI Could Reduce the Impact of Biased Algorithms." *Ventura Beat*.

Clozel, L. (2017, June 30). "Is Your AI Racist? This Lawmaker Wants to Know." *American Banker*.

Cokins, G. (2017, March 22). "Opinion Could IBM's New Deep Learning Service Tool Help Save IT Jobs?" *Information Management*.

Collins, T. (2017, December 18). "Google and Amazon Really DO Want to Spy on You: Patent Reveals Future Version of Their Voice Assistants Will Record Your Conversations to Sell You Products." *Daily Mail*.

Crespo, M. (2017, July 31). "The Future of Work in the Era of Artificial Intelligence." *Equal Times*.

Crosman, P. (2017, August 17). "Why Cybercriminals Like AI As Much As Cyberdefenders Do." *American Banker*.

Daugherty, P. R., & J. Wilson. (2018). *Human + Machine: Reimagining Work in the Age of AI*. Boston, MA: Business Review Press.

Desjardins, J. (2017, August 21). "Visualizing the Massive $15.7 Trillion Impact of AI." *Visual Capitalist*.

de Vos, B. (2018, July 11). "Opinion: These 3 Business Functions Will Be the First to Benefit from Artificial Intelligence." *Information Management*.

DiCamillo, N. (2018, July 12). "Morgan Stanley Draws from 'Hundreds of Conversations' with Experts to Build Its AI." *American Banker*.

Dickson, B. (2017, July 28). "What Is the Future of Artificial Intelligence?" *Tech Talk*.

Donahue, L. "A Primer on Using Artificial Intelligence in the Legal Profession." *Jolt Digest*, January 3, 2018.

Dormehl, L. (2017). *Thinking Machines: The Quest for Artificial Intelligence—and Where It's Taking Us Next*. New York, NY: TarcherPerigee.

Eadicicco, L. (2017, October 13). "Google Searches for Its Voice." *Time for Kids*.

Editors. (2018, July 12). "Smart Solutions Can Help ASEAN Cities Improve Quality-of-Life Indicators by 10−30%." *eGov Innovation*.

Egan, M. (2015, May 13). "Robots Threaten These 8 Jobs." **CNNMoney.com**.

Ekster, G. (2015). Driving Investment Performance with Alternative Data. **integrity-research.com/wp-content/uploads/2015/11/Driving-Investment-Performance-With-Alternative-Data.pdf** (accessed July 2018).

Elgan, M. (2017, April 29). "How the Amazon Echo Look Improves Privacy?" *Computer World*.

Elson, R. J., & LeClerc, R. (2005). Security and Privacy Concerns in the Data Warehouse Environment. *Business Intelligence Journal, 10*(3), 51.

Gaudin, S. (2016, October 26). "1-800-Flowers Wants to Transform Its Business with A.I." *Computer World*.

Gershgorn, D. (2017, October 22). "Google Is Teaching Its AI How Humans Hug, Cook and Fight." *Quartz*. **qz.com/1108090/google-is-teaching-its-ai-how-humans-hug-cook-and-fight/** (accessed April 2018).

Goldman, S. (2018, March 22). "The Ethics of Legal Analytics." **Law.com**.

Guha, A. (2017, June 5). "Labour and Artificial Intelligence: Visions of Despair, Hope and Liberation." **Hindustan Times.com**.

Himmelreich, J. (2018, March 27). "The Ethical Challenges Self-Driving Cars Will Face Every Day." *Smithsonian*.

Hu, F. (2016). *Security and Privacy in Internet of Things (IoTs): Models, Algorithms, and Implementations*. Boca Raton, FL: CRC Press.

Huff, E. (2017, January 17). "Proof That Amazon Devices Are Spies in Your Own Home: Alexa Automatically Orders Product after 'Hearing' Audio in Private Homes." *Natural News*.

Kahn, J. (2017, November 29). "Legal AI Gains Traction as U.K. Startup Targets U.S." *Bloomberg Technology*.

Kaplan, J. (2017). Startup Targets. *Artificial Intelligence: What Everyone Needs to Know*. London, United Kingdom: Oxford University Press.

Kassner, M. (2017, January 2). "5 Ethics Principles Big Data Analysts Must Follow." *Tech Republic*.

Keenan, J. (2018, February 13). "1-800-Flowers.com Using Technology to Win Customers' Hearts This Valentine's Day." *Total Retail*.

Kelly, H. (2018, January 29). "Robots Could Kill Many Las Vegas Jobs." **Money.CNN.com**.

Kiron, D. (2017, January 25). "What Managers Need to Know About Artificial Intelligence." *MIT Sloan Management Review*.

Knight, W. (2018, March 7). "Inside the Chinese Lab That Plans to Rewire the World with AI." *MIT Technology Review*.

Kokalitcheva, K. (2017, May 9). "The Full History of the Uber-Waymo Legal Fight." *Axio*.

Konomi, S., & G. Roussos (ed.). (2016). *Enriching Urban Spaces with Ambient Computing, the Internet of Things, and Smart City Design (Advances in Human and Social Aspects of Technology)*. Hershey, PA: GI Global.

Korolov, M. (2016, December 2). "There Will Still Be Plenty of Work to Go Around So Job Prospects Should Remain Good." *IT World*.

Kottasova, I. (2018, April 12). "Experts Warn Europe: Don't Grant Rights." **Money.CNN.com**.

Kovach, S. (2018, January). "Amazon Has Created a New Computing Platform That Will Future-Proof Your Home." *Business Insider*. **businessinsider.com/amazon-alexa-best-way-future-proof-smart-home-2018-1/** (Accessed July 2018).

Krauth, O. (2018, January 23). "Robot Gender Gap: Women Will Lose More Jobs Due to Automation Than Men, WEF Finds." *Tech Republic*.

Krigsman, M. (2017, January 30). "Artificial Intelligence: Legal, Ethical, and Policy Issues." *ZDNet*.

Kurzer, R. (2017, December 21). "What Is the Future of Artificial Intelligence?" *Martechnology Today*.

Lashinsky, A. (2018, June 21). "Alibaba v. Tencent: The Battle for Supremacy in China." *Fortune*.

Lawson, K. (2017, May 2). "Do You Need a Chief Artificial Intelligence Officer?" *Information Management*.

Leggatt, H. (2017, June 7). "Biggest Stressor in U.S. Workplace Is Fear of Losing Jobs to AI, New Tech." *Biz Report*.

Lev-Ram, M. (2017, September 26). "Tech's Magic 8 Ball Says Embrace the Future." *Fortune*.

Maguire, J. (2017, February 3). "Artificial Intelligence: When Will the Robots Rebel?" *Datamation*. **datamation.com/data-center/artificial-intelligence-when-will-the-robots-rebel.html** (accessed April 2018).

Manyika, J. (2017, May). "Technology, Jobs, and the Future of Work." *McKinsey Global Institute*.

Manyika, J., M. Chi, M. Miremadi, J. Bughin, K. George, P. Willmott, & M. Dewhurst. (2017, January). "Harnessing Automation for a Future That Works." *Report from the McKinsey Global Institute*. **mckinsey.com/global-themes/digital-disruption/harnessing-automation-for-a-future-that-works/** (accessed April 2018).

Marr, B. (2018, June 4). "Artificial Intelligence (AI) in China: The Amazing Ways Tencent Is Driving Its Adoption." *Forbes*.

Marshall, A., & A. Davies. (2018, February 9). "The End of Waymo v. Uber Marks a New Era for Self-Driving Cars: Reality." *Wired*.

Mason, R., F. Mason, & M. Culnan. (1995). *Ethics of Information Management*. Thousand Oaks, CA: Sage.

McCracken, H. (2017, October 10). "How to Stop Worrying and Love the Great AI War of 2018." *Fast Company*.

McFarland, M. (2017a, April 28). "Robots Hit the Streets—and the Streets Hit Back." *CNN Tech*.

McFarland, M. (2017b, September 15). "Robots: Is Your Job At Risk?" CNN News.

Morgan, B. (2017, June 13). "Ethics and Artificial Intelligence with IBM Watson's Rob High." *Forbes*.

Morris, D. (2017, February 18). "Bill Gates Says Robots Should Be Taxed Like Workers." **Fortune.com**.

Newman, D. (2018, January 16). "Top 18 Tech Trends at CES 2018." **Forbes.com**.

Olavsrud, T. (2018, March 15). "6 Data Analytics Trends That Will Dominate 2018." *CIO*.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown Publishing).

Pakzad, R. (2018, January 21). "Ethics in Machine Learning." **Medium.com**.

Palmer, S. (2017, February 26). "The 5 Jobs Robots Will Take First." *Shelly Palmer*.

Perez, A. (2017, May 31). "Opinion Will AI and Machine Learning Replace the Data Scientist?" *Information Management*.

Pham, S. (2018, February 21). "Control AI Now or Brace for Nightmare Future, Experts Warn." **Money.cnn.com** *(News)*.

Press, G. (2017, November 9). "10 Predictions for AI, Big Data, and Analytics in 2018." **Forbes.com**.

Provazza, A. (2017, May 26). "Artificial Intelligence Data Privacy Issues on the Rise." *Tech Target (News)*.

Rainie, L., & J. Anderson. (2017, June 6). "The Internet of Things Connectivity Binge: What Are the Implications?" *Pew Research Center*.

Ransbotham, S. (2016). "How Will Cognitive Technologies Affect Your Organization?" **sloanreview.mit.edu/article/how-will-cognitive-technologies-affect-your-organization/** (accessed July 2018).

Rao, A., J. Voyles, & P. Ramchandani. (2017, December 5). "Top 10 Artificial Intelligence (AI) Technology Trends for 2018." *USBlogs PwC*.

Rayo, E. A. "AI in Law and Legal Practice – A Comprehensive View of 35 Current Applications." *Techemergence*, September 19, 2018.

Rikert, T. (2017, September 25). "Using AI and Machine Learning to Beat the Competition." *NextWorld*. **insights.nextworldcap.com/ai-machine-learning-b01946a089b2** (accessed July 2018).

Ross, J. (2017, July 14). "The Fundamental Flaw in AI Implementation." *MIT Sloan Management Review*. **sloanreview.mit.edu/article/the-fundamental-flaw-in-ai-implementation/** (accessed July 2018).

Sage, A. et al. (2018, February 9). "Waymo Accepts $245 Million and Uber's 'Regret' to Settle Self-Driving Car Dispute." *Reuters (Business News)*.

SAS. (n.d.). "Customer Loyalty Blossoms with Analytics." *SAS Publication*, **sas.com/en_us/customers/1-800-flowers.html/** (accessed July 2018).

SAS. (2018). "Artificial Intelligence for Executives." *White Paper*.

Sharma, K. (2017, June 28). "5 Principles to Make Sure Businesses Design Responsible AI." *Fast Company*.

Shchutskaya, V. (2017, March 20). "3 Major Problems of Artificial Intelligence Implementation into Commercial Projects." *InData Labs*. **https://indatalabs.com/blog/data-science/problems-of-artificial-intelligence-implementation/** (accessed April 2018).

Sherman, E. (2015, February 25). "5 White-Collar Jobs Robots Already Have Taken." **Fortune.com fortune.com/2015/02/25/5-jobs-that-robots-already-are-taking** (accessed April 2018).

Sherman, J. (2018, October 16). "Human-Centered Design for Empathy Values and AI." *AlMed*.

Singh, G. (2017a, September 20). "Opinion: 5 Components That Artificial Intelligence Must Have to Succeed." *Health DataManagement*.

Singh, S. (2017b, December 13). "By 2020, Artificial Intelligence Will Create More Jobs Than It Eliminates: Gartner." *The Economic Times (India)*.

Smith, Ms. (2018, March 12). "Ransomware: Coming to a Robot Near You Soon?" *CSO*, News.

Smith, N. (2018, January 3). "Top 10 Trends for Analytics in 2018." *CIO Knowledge*.

Snyder, A. (2017, September 6). "Executives Say AI Will Change Business, But Aren't Doing Much About It." **Axios.com**.

Sommer, D. (2017, December 20). "Opinion Predictions 2018: 11 Top Trends Driving Business Intelligence." *Information Management*.

Spangler, T. (2017, November 24). "Self-Driving Cars Programmed to Decide Who Dies in a Crash." *USA Today*.

Standage, T. (2016) "The Return of the Machinery Question." Special Report. *The Economist*. **economist.com/sites/default/files/ai_mailout.pdf** (accessed July 2018).

Steinberg, J. (2017, April 26). "Echo Lock: Amazon's New Alexa Device Provide Fashion Advice." *INC*.

Straus, R. (2014, May 31). "Will You Be Replaced by a Robot? We Reveal the 100 Occupations Judged Most and Least at Risk of Automation." **ThisisMoney.com**. **thisismoney.co.uk/money/news/article-2642880/Table-700-jobs-reveals-professions-likely-replaced-robots.html** (accessed April 2018).

Sykes, N. (2018a, March 27). "Opinion: Edge Computing and the Future of the Data Center." *Information Management*.

Sykes, N. (2018b, January 17). "Opinion: 9 Top Trends Impacting the Data Center in 2018." *Information Management*.

Thusoo, A. (2017, September 27). "Opinion: AI Is Changing the Skills Needed of Tomorrow's Data." *Information Management*.

Uzialko, A. (2017, October 13). "AI Comes to Work: How Artificial Intelligence Will Transform Business." *Business News Daily*.

Vanian, J. (2017, July 26). "Mark Zuckerberg Argues Against Elon Musk's View of Artificial Intelligence. . . Again." *Fortune*.

Violino, B. (2017, June 27). "Artificial Intelligence Has Potential to Drive Large Profits." *Information Management*.

Violino, B. (2018, February 21). "Most Workers See Smart Robots As Aid to Their Jobs, Not Threat." *Information Management*.

WallStreetJournal.com. (2016). "What They Know." **wsj.com/public/page/what-they-know-digital-privacy.html** (accessed April 2018).

Welch, D. (2016, July 12). The Battle for Smart Car Data. *Bloomberg Technology*. **bloomberg.com/news/articles/2016-07-12/your-car-s-been-studying-you-closely-and-everyone-wants-the-data** (accessed April 2018).

Weldon, D. (2017a, May 5). "AI Seen as Great 'Equalizer' in Bringing Services to the Masses." *Information Management*.

Weldon, D. (2017a, August 11). "Majority of Workers Welcome Job Impacts of AI, Automation." *Information Management*.

Weldon, D. (2017c, July 21). "Smarter Use of Analytics Offers Top Competitive Advantage." *Information Management*.

Weldon, D. (2018, February 28). "Knowing When It's Time to Appoint a Chief Data Officer." *Information Management*.

Weldon, D. (2018, October 18) "Gartner's top 10 strategic technology trends for 2019." *Information Management*.

West, D. (2018). *The Future of Work: Robots, AI, and Automation*. Washington, DC: Brooking Institute Press.

Wilson, H. et al. (2017, March 23). "The Jobs That Artificial Intelligence Will Create." *MIT Sloan Management Review*.

Yudkowsky (2016, May 5) **youtube.com/watch?v=EUjc1WuyPT8**.

GLOSSARY

active data warehousing *See* real-time data warehousing.

ad hoc query A query that cannot be determined prior to the moment the query is issued.

agency The degree of autonomy vested in a software agent.

Alexa The virtual personal assistant of **Amazon.com**.

algorithm A step-by-step search in which improvement is made at every step until the best solution is found.

ambient computing Electronic environment that is sensitive and responsive to people. The technology serves the environment and acts to support the involved people in their tasks.

analytic hierarchy process (AHP) A modeling structure for representing *multi-criteria* (multiple goals, multiple objectives) *problems*—with sets of criteria and alternatives (choices)—commonly found in business environments.

analytical models Mathematical models into which data are loaded for analysis.

analytical techniques Methods that use mathematical formulas to derive an optimal solution directly or to predict a certain result, mainly in solving structured problems.

analytics The science of analysis.

analytics ecosystem A classification of sectors, technology/solution providers, and industry participants for analytics.

application service provider (ASP) A software vendor that offers leased software applications to organizations.

Apriori algorithm The most commonly used algorithm to discover association rules by recursively identifying frequent itemsets.

area under the ROC curve A graphical assessment technique for binary classification models where the true positive rate is plotted on the *Y*-axis and the false positive rate is plotted on the *X*-axis.

artificial brain People-made machine that attempts to be intelligent, creative, and self-aware.

artificial intelligence (AI) Behavior by a machine that, if performed by a human being, would be called intelligent.

artificial neural network (ANN) Computer technology that attempts to build computers that operate like a human brain. The machines possess simultaneous memory storage and work with ambiguous information. Sometimes called, simply, a *neural network*. *See* neural computing.

association A category of data mining algorithm that establishes relationships about items that occur together in a given record.

asynchronous Occurring at different times.

augmented intelligence This is an alternative conceptualization of artificial intelligence that focuses on AI's assistive role, emphasizing the fact that it is designed to enhance human intelligence rather than replace it.

augmented reality The integration of users' senses with the surrounding environment and information technology. It provides people with real-world interactive experiences with the environment.

authoritative pages Web pages that are identified as particularly popular based on links by other Web pages and directories.

automated decision system (ADS) A business rule–based system that uses intelligence to recommend solutions to repetitive decisions (such as pricing).

automation The process by which special purpose machines or systems are able to complete tasks without human intervention.

autonomous cars A vehicle that can guide itself without human intervention.

autonomous vehicles Self-driving vehicles that do not need a driver and are preprogrammed to drive to destinations; also referred to as robot-driven cars, self-driving cars, and autonomous cars.

autonomy The ability to make your own decisions.

axon An outgoing connection (i.e., terminal) from a biological neuron.

backpropagation The best-known learning algorithm in neural computing where the learning is done by comparing computed outputs to desired outputs of training cases.

backward chaining A search technique (based on if-then rules) used in production systems that begins with the action clause of a rule and works backward through a chain of rules in an attempt to find a verifiable set of condition clauses.

bagging The simplest and most common type of ensemble method; it builds multiple prediction models (e.g., decision trees) from bootstrapped/resampled data and combines the predicted values through averaging or voting.

balanced scorecard (BSC) A performance measurement and management methodology that helps translate an organization's financial, customer, internal process, and learning and growth objectives and targets into a set of actionable initiatives.

Bayes theorem (also called Bayes rule) Named after the British mathematician Thomas Bayes (1701–1761), this is a mathematical formula for determining conditional probabilities.

Bayesian belief networks (or Bayesian networks) These are powerful tools for representing dependency structure among variables in a graphical, explicit, and intuitive way.

Bayesian network model This is a directed acyclic graph where the nodes correspond to the variables, and the arcs

signify conditional dependencies between variables and their possible values.

best practices In an organization, the best methods for solving problems. These are often stored in the knowledge repository of a knowledge management system.

Big Data Data that are characterized by the volume, variety, and velocity that exceed the reach of commonly used hardware environments and/or capabilities of software tools to process.

Big Data analytics Application of analytics methods and tools to Big Data.

boosting This is an ensemble method where a series of prediction models are built progressively to improve the predictive performance of the cases/samples incorrectly predicted by the previous ones.

bootstrapping A sampling technique where a fixed number of instances from the original data is sampled (with replacement) for training and the rest of the data set is used for testing.

bot An intelligent software agent. Bot is an abbreviation of robot and is usually used as part of another term, such as knowbot, softbot, or shopbot.

brainstorming A process for generating creative ideas.

business (or system) analyst An individual whose job is to analyze business processes and the support they receive (or need) from information technology.

business analytics (BA) The application of models directly to business data. Business analytics involve using DSS tools, especially models, in assisting decision makers. *See also* business intelligence (BI).

business intelligence (BI) A conceptual framework for managerial decision support. It combines architecture, databases (or data warehouses), analytical tools, and applications.

business network A group of people who have some kind of commercial relationship; for example, sellers and buyers, buyers among themselves, buyers and suppliers, and colleagues and other colleagues.

business performance management (BPM) An advanced performance measurement and analysis approach that embraces planning and strategy.

business process reengineering (BPR) A methodology for introducing a fundamental change in specific business processes. BPR is usually supported by an information system.

Caffe This is an open-source deep learning framework developed at UC Berkeley and Berkeley AI Research.

case-based reasoning (CBR) A methodology in which knowledge or inferences are derived from historical cases.

categorical data Data that represent the labels of multiple classes used to divide a variable into specific groups.

certainty The business situation where complete knowledge is available so that the decision maker knows exactly what the outcome of each course of action will be.

certainty factors A popular technique for representing uncertainty in expert systems where the belief in an event (or a fact or a hypothesis) is expressed using the expert's unique assessment.

chatbot A robot that a person can chat with (in a text or voice) and get information and advice in natural language.

choice phase A phase where the actual decision and the commitment to follow a certain course of action are made.

chromosome A candidate solution for a genetic algorithm.

classification Supervised induction used to analyze the historical data stored in a database and to automatically generate a model that can predict future behavior.

clickstream analysis The analysis of data that occur in the Web environment.

clickstream data Data that provide a trail of the user's activities and show the user's browsing patterns (e.g., which sites are visited, which pages, how long).

cloud computing Information technology infrastructure (hardware, software, applications, platform) that is available as a service, usually as virtualized resources.

clustering Partitioning a database into segments in which the members of a segment share similar qualities.

cognitive computing The application of knowledge derived from cognitive science in order to simulate the human thought process so that computers can exhibit or support decision-making and problem-solving capabilities.

cognitive limits The limitations of the human mind related to processing information.

cognitive search A new generation of search method that uses artificial intelligence (e.g., advanced indexing, NLP, and machine learning) to return results that are much more relevant to the user.

collaboration hub The central point of control for an e-market. A single collaboration hub (c-hub), representing one e-market owner, can host multiple collaboration spaces (c-spaces) in which trading partners use c-enablers to exchange data with the c-hub.

collaborative filtering A method for generating recommendations from user profiles. It uses preferences of other users with similar behavior to predict the preferences of a particular user.

collaborative planning, forecasting, and replenishment (CPFR) A project in which suppliers and retailers collaborate in their planning and demand forecasting to optimize the flow of materials along the supply chain.

collaborative workspace Is where people can work together from any location at the same or a different time.

collective intelligence The total intelligence of a group. It is also referred to as the *wisdom of the crowd*.

community of practice (COP) A group of people in an organization with a common professional interest, often

self-organized, for managing knowledge in a knowledge management system.

complexity A measure of how difficult a problem is in terms of its formulation for optimization, its required optimization effort, or its stochastic nature.

computer ethics Ethical behavior of people toward information systems and computers in general.

computer vision Computer program that helps to recognize scenery (photos, videos).

confidence In association rules, the conditional probability of finding the RHS of the rule present in a list of transactions where the LHS of the rule already exists.

connection weight The weight associated with each link in a neural network model. Neural networks learning algorithms assess connection weights.

consultation environment The part of an expert system that a nonexpert uses to obtain expert knowledge and advice. It includes the workplace, inference engine, explanation facility, recommended action, and user interface.

content management system (CMS) An electronic document management system that produces dynamic versions of documents and automatically maintains the current set for use at the enterprise level.

content-based filtering A type of filtering that recommends items for a user based on the description of previously evaluated items and information available from the content (e.g., keywords).

convolution In convolutional neural networks, this is a linear operation that aims at extracting simple patterns from sophisticated data patterns.

convolution function This is a parameter sharing method to address the issue of computational efficiency in defining and training a very large number of weight parameters that exist in CNN.

convolution layer This is a layer containing a convolution function in a CNN.

convolutional neural networks (CNNs) These are among the most popular deep learning methods. CNNs are in essence a variation of the deep MLP-type neural network architecture, initially designed for computer vision applications (e.g., image processing, video processing, text recognition) but also applicable to nonimage data sets.

corporate (enterprise) portal A gateway for entering a corporate Web site. A corporate portal enables communication, collaboration, and access to company information.

corpus In linguistics, a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.

CRISP-DM A cross-industry standardized process of conducting data mining projects, which is a sequence of six steps that starts with a good understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfied the specific business need.

critical event processing A method of capturing, tracking, and analyzing streams of data to detect certain events (out of normal happenings) that are worthy of the effort.

critical success factors (CSF) Key factors that delineate the areas that an organization must excel at to be successful in its market space.

crowdsourcing Outsourcing tasks (work) to a large group of people.

cube A subset of highly interrelated data that is organized to allow users to combine any attributes in a cube (e.g., stores, products, customers, suppliers) with any metrics in the cube (e.g., sales, profit, units, age) to create various two-dimensional views, or *slices*, that can be displayed on a computer screen.

customer experience management (CEM) Applications designed to report on the overall user experience by detecting Web application issues and problems, by tracking and resolving business process and usability obstacles, by reporting on site performance and availability, by enabling real-time alerting and monitoring, and by supporting deep diagnosis of observed visitor behavior.

dashboard A visual presentation of critical data for executives to view. It allows executives to see hot spots in seconds and explore the situation.

data Raw facts that are meaningless by themselves (e.g., names, numbers).

data cube A two-dimensional, three-dimensional, or higher-dimensional object in which each dimension of the data represents a measure of interest.

data integration Integration that comprises three major processes: data access, data federation, and change capture. When these three processes are correctly implemented, data can be accessed and made accessible to an array of ETL, analysis tools, and data warehousing environments.

data integrity A part of data quality where the accuracy of the data (as a whole) is maintained during any operation (such as transfer, storage, or retrieval).

data mart A departmental data warehouse that stores only relevant data.

data mining A process that uses statistical, mathematical, artificial intelligence, and machine-learning techniques to extract and identify useful information and subsequent knowledge from large databases.

data quality (DQ) The holistic quality of data, including their accuracy, precision, completeness, and relevance.

data scientist A person employed to analyze and interpret complex digital data to assist a business in decision making.

data stream mining The process of extracting novel patterns and knowledge structures from continuously streaming data records. *See* stream analytics.

data visualization A graphical, animation, or video presentation of data and the results of data analysis.

data warehouse (DW) A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format.

database A collection of files that are viewed as a single storage concept. The data are then available to a wide range of users.

database management system (DBMS) Software for establishing, updating, and querying (e.g., managing) a database.

deception detection A way of identifying deception (intentionally propagating beliefs that are not true) in voice, text, and/or body language of humans.

decision analysis A modeling approach that deals with decision situations that involve a finite and usually not too large number of alternatives.

decision making The action of selecting among alternatives.

decision or normative analytics Also called prescriptive analytics, this is a type of analytics modeling that aims at identifying the best possible decision from a large set of alternatives.

decision room Expensive, customized, special-purpose facility with a group support system in which PCs are available to some or all participants. The objective is to enhance group work.

decision support systems (DSS) A conceptual framework for a process of supporting managerial decision making, usually by modeling problems and employing quantitative models for solution analysis.

decision table A tabular representation of possible condition combinations and outcomes.

decision tree A graphical presentation of a sequence of interrelated decisions to be made under assumed risk. This technique classifies specific entities into particular classes based upon the features of the entities; a root is followed by internal nodes, each node (including root) is labeled with a question, and arcs associated with each node cover all possible responses.

decision variable The variable of interest.

deep learning The newest and perhaps the most popular member of the artificial intelligence and machine learning family, deep learning has a goal similar to those of the other machine learning methods that came before it: mimic the thought process of humans—using mathematical algorithms to learn from data (both representation of the variables and their interrelationships).

deep neural networks These are a part of deep learning algorithms where numerous hidden layers of neurons are used to capture the complex relationships from very large training data sets.

defuzzification The process of creating a crisp solution from a fuzzy logic solution.

dendrite The part of a biological neuron that provides inputs to the cell.

dependent data mart A subset that is created directly from a data warehouse.

descriptive (or reporting) analytics An earlier phase in analytics continuum that deals with describing the data answering the questions of what happened and why did it happen.

design phase This phase involves inventing, developing, and analyzing possible courses of action.

development environment The part of an expert system that a builder uses. It includes the knowledge base and the inference engine, and it involves knowledge acquisition and improvement of reasoning capability. The knowledge engineer and the expert are considered part of the environment.

dimensional modeling A retrieval-based system that supports high-volume query access.

directory A catalog of all the data in a database or all the models in a model base.

discrete event simulation A type of simulation modeling where a system is studied based on the occurrence of events/interaction between different parts (entities/resources) of the system.

distance measure A method used to calculate the closeness between pairs of items in most cluster analysis methods. Popular distance measures include Euclidean distance (the ordinary distance between two points that one would measure with a ruler) and Manhattan distance (also called the rectilinear distance, or taxicab distance, between two points).

distributed artificial intelligence (DAI) A multiple-agent system for problem solving. DAI involves splitting a problem into multiple cooperating systems to derive a solution.

DMAIC A closed-loop business improvement model that includes these steps: defining, measuring, analyzing, improving, and controlling a process.

document management systems (DMS) Information systems (e.g., hardware, software) that allow the flow, storage, retrieval, and use of digitized documents.

drill-down The investigation of information in detail (e.g., finding not only total sales but also sales by region, by product, or by salesperson). Finding the detailed sources.

DSS application A DSS program built for a specific purpose (e.g., a scheduling system for a specific company).

dynamic models A modeling technique to capture/study systems that evolve over time.

Echo The speaker that works together with Alexa.

effectiveness The degree of goal attainment. Doing the right things.

effectors An effector is a device designed for robots to interact with the environment.

efficiency The ratio of output to input. Appropriate use of resources. Doing things right.

electronic brainstorming A computer-supported methodology of idea generation by association. This group process uses analogy and synergy.

electronic meeting systems (EMS) An information technology–based environment that supports group meetings (groupware), which may be distributed geographically and temporally.

ensembles (or more appropriately called model ensembles or ensemble modeling) These are combinations of the outcomes produced by two or more analytics models into a compound output. Ensembles are primarily used for prediction modeling where the scores of two or more models are combined to produce a better prediction.

Enterprise 2.0 Technologies and business practices that free the workforce from the constraints of legacy communication and productivity tools such as e-mail. Provides business managers with access to the right information at the right time through a Web of interconnected applications, services, and devices.

enterprise application integration (EAI) A technology that provides a vehicle for pushing data from source systems into a data warehouse.

enterprise data warehouse (EDW) An organizational-level data warehouse developed for analytical purposes.

entropy A metric that measures the extent of uncertainty or randomness in a data set. If all the data in a subset belong to just one class, then there is no uncertainty or randomness in that data set, and therefore the entropy is zero.

environmental scanning and analysis A continuous process of intelligence building identification of problems and/or opportunities via acquisition and analysis of data/information.

evolutionary algorithm A class of heuristic-based optimization algorithms modeled after the natural process of biological evolution, such as genetic algorithms and genetic programming.

expert A human being who has developed a high level of proficiency in making judgments in a specific, usually narrow, domain.

expert location system An interactive computerized system that helps employees find and connect with colleagues who have expertise required for specific problems—whether they are across the county or across the room—in order to solve specific, critical business problems in seconds.

expert system (ES) shell A computer program that facilitates relatively easy implementation of a specific expert system. Analogous to a DSS generator.

expert systems Computerized systems that transfer expert and documented knowledge to machines that help nonexperts use this knowledge for decision making.

expertise The set of capabilities that underlines the performance of human experts, including extensive domain knowledge, heuristic rules that simplify and improve approaches to problem solving, metaknowledge and metacognition, and collective forms of behavior that afford great economy in a skilled performance.

explanation subsystem The component of an expert system that can explain the system's reasoning and justify its conclusions.

explicit knowledge Knowledge that deals with objective, rational, and technical material (e.g., data, policies, procedures, software, documents). Also known as *leaky knowledge*.

extraction The process of capturing data from several sources, synthesizing them, summarizing them, determining which of them are relevant, and organizing them, resulting in their effective integration.

facilitator (in a GSS) A person who plans, organizes, and electronically controls a group in a collaborative computing environment.

forecasting Using the data from the past to foresee the future values of a variable of interest.

forward chaining A data-driven search in a rule-based system.

functional integration The provision of different support functions as a single system through a single, consistent interface.

fuzzification A process that converts an accurate number into a fuzzy description, such as converting from an exact age into categories such as young and old.

fuzzy logic A logically consistent way of reasoning that can cope with uncertain or partial information. Fuzzy logic is characteristic of human thinking and expert systems.

fuzzy set A set theory approach in which set membership is less precise than having objects strictly in or out of the set.

genetic algorithm A software program that learns in an evolutionary manner, similar to the way biological systems evolve.

geographic information systems (GIS) An information system capable of integrating, editing, analyzing, sharing, and displaying geographically referenced information.

Gini index A metric that is used in economics to measure the diversity of the population. The same concept can be used to determine the purity of a specific class as a result of a decision to branch along a particular attribute/variable.

global positioning systems (GPS) Wireless devices that use satellites to enable users to detect the position on earth of items (e.g., cars or people) the devices are attached to, with reasonable precision.

goal seeking A prescriptive analytics method where first a goal (a target/desired value) is set, and then the satisfying set of input variable values is identified.

Google Assistant An upcoming virtual personal assistant for use in several of Google's products.

grain A definition of the highest level of detail that is supported in a data warehouse.

graphic processing unit (GPU) It is the part of a computer that normally processes/renders graphical outputs; nowadays, it is also being used for efficient processing of deep learning algorithms.

graphical user interface (GUI) An interactive, user-friendly interface in which, by using icons and similar objects, the user can control communication with a computer.

group decision making A situation in which people make decisions together.

group decision support system (GDSS) An interactive computer-based system that facilitates the solution of semistructured and unstructured problems by a group of decision makers.

group support system (GSS) Information system, specifically DSS, that supports the collaborative work of groups.

group work Any work being performed by more than one person.

groupthink Continual reinforcement of an idea by group members in a meeting.

groupware Computerized technologies and methods that aim to support people working in groups.

groupwork Any work being performed by more than one person.

Hadoop An open-source framework for processing, storing, and analyzing massive amounts of distributed, unstructured data.

Hadoop Distributed File System (HDFS) A distributed file management system that lends itself well to processing large volumes of unstructured data (i.e., Big Data).

heterogeneous ensembles These combine the outcomes of two or more different types of models such as decision trees, artificial neural networks, logistic regression, support vector machines, and others.

heuristic programming The use of heuristics in problem solving.

heuristics Informal, judgmental knowledge of an application area that constitutes the rules of good judgment in the field. Heuristics also encompasses the knowledge of how to solve problems efficiently and effectively, how to plan steps in solving a complex problem, how to improve performance, and so forth.

hidden layer The middle layer of an artificial neural network that has three or more layers.

Hive Hadoop-based data warehousing like framework originally developed by Facebook.

homogeneous ensembles combine the outcomes of two or more of the same type of models such as decision trees.

hub One or more Web pages that provide a collection of links to authoritative pages.

hybrid (integrated) computer system Different but integrated computer support systems used together in one decision-making situation.

hyperlink-induced topic search (HITS) The most popular publicly known and referenced algorithm in Web mining used to discover hubs and authorities.

hyperplane A geometric concept commonly used to describe the separation surface between different classes of things within a multidimensional space.

hypothesis-driven data mining A form of data mining that begins with a proposition by the user, who then seeks to validate the truthfulness of the proposition.

IBM Watson It is an extraordinary computer system—a novel combination of advanced hardware, software, and machine-learning algorithms—designed to answer questions posed in natural human language.

IBM SPSS Modeler A very popular, commercially available, comprehensive data, text, and Web mining software suite developed by SPSS (formerly Clementine).

idea generation The process by which people generate ideas, usually supported by software (e.g., developing alternative solutions to a problem). Also known as *brainstorming*.

ImageNet This is an ongoing research project that provides researchers with a large database of images, each linked to a set of synonym words (known as synset) from WordNet (a word hierarchy database).

implementation phase A phase that involves putting a recommended solution to work, not necessarily implementing a computer system.

inference engine The part of an expert system that actually performs the reasoning function.

influence diagram A graphical representation of a given mathematical model.

information Data organized in a meaningful way.

information fusion (or simply, fusion) A type of heterogeneous model ensembles that combines different types of prediction models using a weighted average, where the weights are determined from the individual models' predictive accuracies.

information gain The splitting mechanism used in ID3 (a popular decision-tree algorithm).

information overload An excessive amount of information being provided, making processing and absorbing tasks very difficult for the individual.

intelligence A degree of reasoning and learned behavior, usually task or problem-solving oriented.

intelligence phase A phase where the decision maker examines reality and identifies and defines the problem.

intelligent agent An autonomous, small computer program that acts upon changing environments as directed by stored knowledge.

intelligent database A database management system exhibiting artificial intelligence features that assist the user or designer; often includes ES and intelligent agents.

interactivity A characteristic of software agents that allows them to interact (communicate and/or collaborate) with each other without having to rely on human intervention.

intermediate result variable A variable used in modeling to identify intermediate outcomes.

Internet of Things (IoT) The technological phenomenon of connecting a variety of devices in the physical world to each other and to the computing systems via the Internet.

Internet of Things ecosystem All components that enable organizations to use IoT; includes the "things," connections, features, procedures, analytics, data, and security.

Internet telephony *See* Voice over IP (VoIP).

interval data Variables that can be measured on interval scales.

inverse document frequency A common and very useful transformation of indices in a term-by-document matrix that reflects both the specificity of words (document frequencies) as well as the overall frequencies of their occurrences (term frequencies).

iterative design A systematic process for system development that is used in management support systems (MSS). Iterative design involves producing a first version of MSS, revising it, producing a second design version, and so on.

Keras An open-source neural network library written in Python that functions as a high-level application programming interface (API) and is able to run on top of various deep learning frameworks including Theano and TensorFlow.

kernel trick In machine learning, a method for using a linear classifier algorithm to solve a nonlinear problem by mapping the original nonlinear observations onto a higher-dimensional space, where the linear classifier is subsequently used; this makes a linear classification in the new space equivalent to a nonlinear classification in the original space.

kernel type In kernel trick, a type of transformation algorithm used to represent data items in a Euclidean space. The most commonly used kernel type is the radial basis function.

key performance indicator (KPI) Measure of performance against a strategic objective and goal.

***k*-fold cross-validation** A popular accuracy assessment technique for prediction models where the complete data set is randomly split into *k* mutually exclusive subsets of approximately equal size. The classification model is trained and tested *k* times. Each time it is trained on all but one fold and then tested on the remaining single fold. The cross-validation estimate of the overall accuracy of a model is calculated by simply averaging the *k* individual accuracy measures.

***k*-nearest neighbor (*k*-NN)** A prediction method for classification as well as regression-type prediction problems where the prediction is made based on the similarity to *k* neighbors.

KNIME An open-source, free-of-charge, platform-agnostic analytics software tool (available at www.knime.org).

knowledge Understanding, awareness, or familiarity acquired through education or experience; anything that has been learned, perceived, discovered, inferred, or understood; the ability to use information. In a knowledge management system, knowledge is information in action.

knowledge acquisition The extraction and formulation of knowledge derived from various sources, especially from experts.

knowledge audit The process of identifying the knowledge an organization has, who has it, and how it flows (or does not) through the enterprise.

knowledge base A collection of facts, rules, and procedures organized into schemas. A knowledge base is the assembly of all the information and knowledge about a specific field of interest.

knowledge discovery in databases (KDD) A machine-learning process that performs rule induction or a related procedure to establish knowledge from large databases.

knowledge engineer An artificial intelligence specialist responsible for the technical side of developing an expert system. The knowledge engineer works closely with the domain expert to capture the expert's knowledge in a knowledge base.

knowledge engineering The engineering discipline in which knowledge is integrated into computer systems to solve complex problems that normally require a high level of human expertise.

knowledge management system (KMS) A system that facilitates knowledge management by ensuring knowledge flow from the person(s) who knows to the person(s) who needs to know throughout the organization; knowledge evolves and grows during the process.

knowledge management The active management of the expertise in an organization. It involves collecting, categorizing, and disseminating knowledge.

knowledge repository The actual storage location of knowledge in a knowledge management system. A knowledge repository is similar in nature to a database but is generally text oriented.

knowledge rules A collection of if-then rules that represents the deep knowledge about a specific problem.

knowledge-based economy The modern, global economy, which is driven by what people and organizations know rather than only by capital and labor. An economy based on intellectual assets.

knowledge-based system (KBS) Typically, a rule-based system for providing expertise. A KBS is identical to an expert system, except that the source of expertise may include documented knowledge.

knowledge-refining system A system that is capable of analyzing its own performance, learning, and improving itself for future consultations.

Kohonen self-organizing feature map (SOM) A type of neural network model for machine learning.

leaky knowledge *See* explicit knowledge.

learning A process of self-improvement where the new knowledge is obtained through a process by using what is already known.

learning algorithm The training procedure used by an artificial neural network.

learning organization An organization that is capable of learning from its past experience, implying the existence of an organizational memory and a means to save, represent, and share it through its personnel.

learning rate A parameter for learning in neural networks. It determines the portion of the existing discrepancy that must be offset.

linear programming (LP) A mathematical modeling technique used to represent and solve constraint optimization problems.

link analysis The linkage among many objects of interest is discovered automatically, such as the link between Web pages and referential relationships among groups of academic publication authors.

literature mining A popular application area for text mining where a large collection of literature (articles, abstracts, book excerpts, and commentaries) in a specific area is processed using semiautomated methods in order to discover novel patterns.

long short-term memory (or LSTM) networks A variation of recurrent neural networks that are known as the most effective sequence modeling techniques and are the foundation of many practical applications.

machine learning Teaching computers to learn from examples and large amounts of data.

machine vision Technology and methods used to provide image-based automated inspection and analysis for applications such as robot guides, process controls, automated vehicles, and inspections.

management science (MS) The application of a scientific approach and mathematical models to the analysis and solution of managerial decision situations (e.g., problems, opportunities). Also known as *operations research* (OR).

management support system (MSS) A system that applies any type of decision support tool or technique to managerial decision making.

MapReduce A technique to distribute the processing of very large multistructured data files across a large cluster of machines.

mathematical (quantitative) model A system of symbols and expressions that represent a real situation.

mathematical programming A family of analytic tools designed to help solve managerial problems in which the decision maker must allocate scarce resources among competing activities to optimize a measurable goal.

mental model The mechanisms or images through which a human mind performs sense-making in decision making.

metadata Data about data. In a data warehouse, metadata describe the contents of a data warehouse and the manner of its use.

middleware Software that links application modules from different computer languages and platforms.

mobile agent An intelligent software agent that moves across different system architectures and platforms or from one Internet site to another, retrieving and sending information.

mobility The degree to which agents travel through a computer network.

model base management system (MBMS) Software for establishing, updating, combining, and so on (e.g., managing) a DSS model base.

model base A collection of preprogrammed quantitative models (e.g., statistical, financial, optimization) organized as a single unit.

model mart A small, generally departmental repository of knowledge created by using knowledge-discovery techniques on past decision instances. Model marts are similar to data marts. *See* model warehouse.

model warehouse A large, generally enterprise-wide repository of knowledge created by using knowledge discovery techniques on past decision instances. Model warehouses are similar to data warehouses. *See* model mart.

momentum A learning parameter in backpropagation neural networks.

Monte Carlo simulation A simulation technique that relies on change/probability distribution to represent the uncertainty in the modeling of the decision problem.

multiagent system A system with multiple cooperating software agents.

multidimensional analysis (modeling) A modeling method that involves data analysis in several dimensions.

multidimensional database A database in which the data are organized specifically to support easy and quick multidimensional analysis.

multidimensional OLAP (MOLAP) OLAP implemented via a specialized multidimensional database (or data store) that summarizes transactions into multidimensional views ahead of time.

multidimensionality The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions).

multiple goals Having more than just one goal to consider in an optimization problem.

mutation A genetic operator that causes a random change in a potential solution.

naïve Bayes A simple probability-based classification method derived from the well-known Bayes' theorem. It is one of the machine-learning techniques applicable to classification-type prediction problems.

narrow (weak) AI A form of AI specifically designed to be focused on a narrow task and to seem very intelligent at it.

natural language processing Technology that allows people to communicate with a computer in their native language.

neural (computing) networks A computer design aimed at building intelligent computers that operate in a manner modeled on the functioning of the human brain.

neural computing An experimental computer design aimed at building intelligent computers that operate in a manner modeled on the functioning of the human brain. *See* artificial neural network (ANN).

neural network *See* artificial neural network (ANN).

neuron A cell (i.e., processing element) of a biological or artificial neural network.

nominal data A type of data that contains measurements of simple codes assigned to objects as labels, which are not measurements. For example, the variable *marital status* can be generally categorized as (1) single, (2) married, and (3) divorced.

nominal group technique (NGT) A simple brainstorming process for nonelectronic meetings.

normative model A model that prescribes how a system should operate.

NoSQL (a.k.s. not only SQL) A new paradigm to store and process large volumes of unstructured, semistructured, and multistructured data.

nucleus The central processing portion of a neuron.

numeric data A type of data that represent the numeric values of specific variables. Examples of numerically valued variables include age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees).

object A person, place, or thing about which information is collected, processed, or stored.

object-oriented model base management system (OOMBMS) An MBMS constructed in an object-oriented environment.

online (electronic) workspace Online screens that allow people to share documents, files, project plans, calendars, and so on in the same online place, though not necessarily at the same time.

online analytical processing (OLAP) An information system that enables the user, while at a PC, to query the system, conduct an analysis, and so on. The result is generated in seconds.

online transaction processing (OLTP) A transaction system that is primarily responsible for capturing and storing data related to day-to-day business functions.

online workspace A place where participants can collaborate while working in different time.

operational data store (ODS) A type of database often used as an interim area for a data warehouse, especially for customer information files.

operational models Models that represent problems for the operational level of management.

operational plan A plan that translates an organization's strategic objectives and goals into a set of well-defined tactics and initiatives, resource requirements, and expected results.

optimal solution The best possible solution to a problem.

optimization The process of identifying the best possible solution to a problem.

ordinal data Data that contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable *credit score* can be generally categorized as (1) low, (2) medium, and (3) high.

organizational agent An agent that executes tasks on behalf of a business process or computer application.

organizational culture The aggregate attitudes in an organization concerning a certain issue (e.g., technology, computers, DSS).

organizational knowledge base An organization's knowledge repository.

organizational learning The process of capturing knowledge and making it available enterprise-wide.

organizational memory That which an organization knows.

ossified case A case that has been analyzed and has no further value.

PageRank A link analysis algorithm, named after Larry Page—one of the two founders of Google as a research project at Stanford University in 1996, and used by the Google Web search engine.

paradigmatic case A case that is unique and that can be maintained to derive new knowledge for the future.

parallel processing An advanced computer processing technique that allows a computer to perform multiple processes at once, in parallel.

parallelism In a group support system, a process gain in which everyone in a group can work simultaneously (e.g., in brainstorming, voting, ranking).

parameter Numeric constants used in mathematical modeling.

part-of-speech tagging The process of marking up the words in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and context of its use.

patents A right granted for exclusive royalty or copyright for novel inventions that would not have been obvious improvements of a known technology.

pattern recognition A technique of matching an external pattern to a pattern stored in a computer's memory (i.e., the process of classifying data into predetermined categories). Pattern recognition is used in inference engines, image processing, neural computing, and speech recognition.

perceptron An early neural network structure that uses no hidden layer.

performance measurement system A system that assists managers in tracking the implementations of business strategy by comparing actual results against strategic goals and objectives.

perpetual analytics An analytics practice that continuously evaluates every incoming data point (i.e., observation) against all prior observations to identify patterns/anomalies.

personal agent An agent that performs tasks on behalf of individual users.

physical integration The seamless integration of several systems into one functioning system.

Pig A Hadoop-based query language developed by Yahoo!.

polysemes Words also called *homonyms*, they are syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., *bow* can mean "to bend forward," "the front of the ship," "the weapon that shoots arrows," or "a kind of tied ribbon").

pooling In CNN, it refers to the process of consolidating the elements in the input matrix in order to produce a smaller output matrix, while maintaining the important features.

portal A gateway to Web sites. Portals can be public (e.g., Yahoo!) or private (e.g., corporate portals).

practice approach An approach toward knowledge management that focuses on building the social environments or communities of practice necessary to facilitate the sharing of tacit understanding.

prediction The act of telling about the future.

predictive analysis Use of tools that help determine the probable future outcome for an event or the likelihood of a situation occurring. These tools also identify relationships and patterns.

predictive analytics A business analytical approach toward forecasting (e.g., demand, problems, opportunities) that is used instead of simply reporting data as they occur.

prescriptive analytics A branch of business analytics that deals with finding the best possible solution alternative for a given problem.

principle of choice The criterion for making a choice among alternatives.

privacy Right to be left alone and to be free from unreasonable personal intrusions.

private agent An agent that works for only one person.

problem ownership The jurisdiction (authority) to solve a problem.

problem solving A process in which one starts from an initial state and proceeds to search through a problem space to identify a desired goal.

process approach An approach to knowledge management that attempts to codify organizational knowledge through formalized controls, processes, and technologies.

process gain In a group support system, improvements in the effectiveness of the activities of a meeting.

process loss In a group support system, degradation in the effectiveness of the activities of a meeting.

processing element (PE) A neuron in a neural network.

production rules The most popular form of knowledge representation for expert systems where atomic pieces of knowledge are represented using simple if-then structures.

prototyping In system development, a strategy in which a scaled-down system or portion of a system is constructed in a short time, tested, and improved in several iterations.

public agent An agent that serves any user.

quantitative model Mathematical models that rely on numeric/quantifiable measures.

quantitative software package A preprogrammed (sometimes called *ready-made*) model or optimization system. These packages sometimes serve as building blocks for other quantitative models.

query facility The (database) mechanism that accepts requests for data, accesses them, manipulates them, and queries them.

radio-frequency identification (RFID) A form of wireless communication between tags (integrated circuit with an antenna) and readers (also called an interrogator) to uniquely identify an object.

random forests First introduced by Breiman (2000) as a modification to the simple bagging algorithm, it uses bootstrapped samples of data and a randomly selected subset of variables to build a number of decision trees, and then combines their output via the simple voting.

rapid application development (RAD) A development methodology that adjusts a system development life cycle so that parts of the system can be developed quickly, thereby enabling users to obtain some functionality as soon as possible. RAD includes methods of phased development, prototyping, and throwaway prototyping.

RapidMiner A popular, open-source, free-of-charge data mining software suite that employs a graphically enhanced user interface, a rather large number of algorithms, and a variety of data visualization features.

ratio data Continuous data where both differences and ratios are interpretable. The distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value.

reality mining Data mining of location-based data.

real-time data warehousing The process of loading and providing data via a data warehouse as they become available.

real-time expert system An expert system designed for online dynamic decision support. It has a strict limit on response time; in other words, the system always produces a response by the time it is needed.

recommendation system (agent) A computer system that can suggest new items to a user based on his or her revealed preference. It may be content based or use collaborative filtering to suggest items that match the preference of the user. An example is **Amazon.com**'s "Customers who bought this item also bought …" feature.

recommendation systems Systems that recommend products and services to individuals based on what they know about the individuals' preferences

recurrent neural networks (RNNs) The type of neural networks that have memory and can apply that memory to determine the future outputs.

regression A data mining method for real-world prediction problems where the predicted values (i.e., the output variable or dependent variable) are numeric (e.g., predicting the temperature for tomorrow as 68°F).

reinforcement learning A sub-area of machine learning that is concerned with learning-by-doing-and-measuring to maximize some notion of long-term reward. Reinforcement learning differs from supervised learning in that correct input/output pairs are never presented to the algorithm.

relational database A database whose records are organized into tables that can be processed by either relational algebra or relational calculus.

relational model base management system (RMBMS) A relational approach (as in relational databases) to the design and development of a model base management system.

relational OLAP (ROLAP) The implementation of an OLAP database on top of an existing relational database.

report Any communication artifact prepared with the specific intention of conveying information in a presentable form.

representation learning A type of machine learning in which the emphasis is on learning and discovering features/variables by the system in addition to mapping of those features to the output/target variable.

reproduction The creation of new generations of improved solutions with the use of a genetic algorithm.

result (outcome) variable A variable that expresses the result of a decision (e.g., one concerning profit), usually one of the goals of a decision-making problem.

revenue management systems Decision-making systems used to make optimal price decisions in order to maximize revenue, based upon previous demand history as well as forecasts of demand at various pricing levels and other considerations.

RFID A generic technology that refers to the use of radio-frequency waves to identify objects.

risk A probabilistic or stochastic decision situation.

risk analysis Use of mathematical modeling to assess the nature of risk (variability) for a decision situation.

robo advisors Virtual personal assistants that contain professional knowledge so they can advise people in several fields, such as in finance and investment.

robot Electromechanical device that is guided by a computer program to perform physical and mental activities.

rule-based system A system in which knowledge is represented completely in terms of rules (e.g., a system based on production rules).

SAS Enterprise Miner A comprehensive, commercial data mining software tool developed by SAS Institute.

satisficing A process by which one seeks a solution that will satisfy a set of constraints. In contrast to optimization, which seeks the best possible solution, satisficing simply seeks a solution that will work well enough.

scenario A statement of assumptions and configurations concerning the operating environment of a particular system at a particular time.

scene recognition Activity performed by a computer vision that enables recognition of objects, scenery, and photos.

scorecard A visual display that is used to chart progress against strategic and tactical goals and targets.

screen sharing Software that enables group members, even in different locations, to work on the same document, which is shown on the PC screen of each participant.

search engine A program that finds and lists Web sites or pages (designated by URLs) that match some user-selected criteria.

search engine optimization (SEO) The intentional activity of affecting the visibility of an e-commerce site or a Web site in a search engine's natural (unpaid or organic) search results.

self-organizing A neural network architecture that uses unsupervised learning.

semantic Web An extension of the current Web, in which information is given well-defined meanings, better enabling computers and people to work in cooperation.

semantic Web services An XML-based technology that allows semantic information to be represented in Web services.

semistructured problem A category of decision problems where the decision process has some structure to it but still requires subjective analysis and an iterative approach.

SEMMA An alternative process for data mining projects proposed by the SAS Institute. The acronym "SEMMA" stands for "sample, explore, modify, model, and assess."

sensitivity analysis A study of the effect of a change in one or more input variables on a proposed solution.

sensitivity analysis simulation The process to investigate the effect of varying a fixed input or a distribution parameter for a simulated input over a specified set of values.

sensor Electronic device that automatically collects data about events or changes in its environment.

sentiment A settled opinion reflective of one's feelings.

sentiment analysis The technique used to detect favorable and unfavorable opinions toward specific products and services using a large number of textual data sources (customer feedback in the form of Web postings).

SentiWordNet An extension of WordNet to be used for sentiment identification. *See* WordNet.

sequence discovery The identification of associations over time.

sequence mining A pattern discovery method where relationships among the things are examined in terms of their order of occurrence to identify associations over time.

shopbot Robot that helps with online shopping by collecting shopping information (search) and conducting price and capability comparisons.

sigmoid (logical activation) function An *S*-shaped transfer function in the range of 0 to 1.

simple split Data are partitioned into two mutually exclusive subsets called a *training set* and a *test set* (or *holdout set*). It is common to designate two-thirds of the data as the training set and the remaining one-third as the test set.

simulation An imitation of reality in computers.

singular value decomposition (SVD) Closely related to principal components analysis, reduces the overall dimensionality of the input matrix (number of input documents by number of extracted terms) to a lower dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents).

Siri Virtual intelligent personal assistant from Apple Computer.

Six Sigma A performance management methodology aimed at reducing the number of defects in a business process to as close to zero defects per million opportunities (DPMO) as possible.

smart appliances Appliances with sensors or smart sensors that occupy smart homes and can be controlled from a distance.

smart cities Cities where many smart things are connected and controlled, including transportation, government services, emergency services, medical services, educational systems, utilities, and possibly homes and public buildings.

smart factory A flexible system that can self-optimize performance across a broader network and self-adapt to and learn from new conditions.

smart homes Homes where the appliances, security, entertainment, and other components are automated, interconnected (frequently wirelessly), and centrally controlled (e.g., via smartphone apps).

smart sensors Sensors with add-on microprocessing capability and possibly other features to best support IoT by processing the collected data.

social analytics The monitoring, analyzing, measuring, and interpreting digital interactions and relationships of people, topics, ideas, and content.

social media The online platforms and tools that people use to share opinions, experiences, insights, perceptions, and various media, including photos, videos, or music, with each other. The enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks.

social media analytics The systematic and scientific way to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness.

social network analysis (SNA) The mapping and measuring of relationships and information flows among people, groups, organizations, computers, and other information- or knowledge-processing entities. The nodes in the network are the people and groups, whereas the links show relationships or flows between the nodes.

social robots An autonomous robot that interacts and communicates with humans or other autonomous physical agents by following social behaviors and rules attached to its role.

software agent A piece of autonomous software that persists to accomplish the task it is designed for (by its owner).

software-as-a-service (SaaS) Software that is rented instead of sold.

Spark An open-source engine developed specifically for handling large-scale data processing for analytics.

speech (voice) understanding Computer systems that attempt to understand words or phrases of human speech, i.e., the natural language spoken by people.

speech analytics A growing field of science that allows users to analyze and extract information from both live and recorded conversations.

stacking (a.k.a. stacked generalization or super learner) A part of heterogeneous ensemble methods where a two-step modeling process is used—first the individual prediction models of different types are built and then a meta-model (a model of the individual models) is built.

staff assistant An individual who acts as an assistant to a manager.

static models A model that captures a snapshot of the system, ignoring its dynamic features.

status report A report that provides the most current information on the status of an item (e.g., orders, expenses, production quantity).

stemming A process of reducing words to their respective root forms in order to better represent them in a text mining project.

stochastic gradient boosting First created by Jerry Friedman at Stanford University in 2001, this is a popular boosting algorithm that uses prediction residuals/errors to guide the gradual development of the future decision trees.

stop words Words that are filtered out prior to or after processing of natural language data (i.e., text).

story A case with rich information and episodes. Lessons may be derived from this kind of case in a case base.

strategic goal A quantified objective that has a designated time period.

strategic models Models that represent problems for the strategic level (i.e., executive level) of management.

strategic objective A broad statement or general course of action that prescribes targeted directions for an organization.

strategic theme A collection of related strategic objectives used to simplify the construction of a strategic map.

strategic vision A picture or mental image of what the organization should look like in the future.

strategy map A visual display that delineates the relationships among the key organizational objectives for all four balanced scorecard perspectives.

stream analytics A term commonly used for extracting actionable information from continuously flowing/streaming data sources.

strong (general) AI A form of AI capable of all and any cognitive functions that a human may have and is in essence no different from a real human mind.

structured problem A decision situation where a specific set of steps can be followed to make a straightforward decision.

structured query language (SQL) A data definition and management language for relational databases. SQL front ends most relational DBMS.

summation function A mechanism to add all the inputs coming into a particular neuron.

supervised learning A method of training artificial neural networks in which sample cases are shown to the network as input, and the weights are adjusted to minimize the error in the outputs.

support The measure of how often products and/or services appear together in the same transaction; that is, the proportion of transactions in the data set that contain all of the products and/or services mentioned in a specific rule.

support vector machines (SVM) A family of generalized linear models, which achieve a classification or regression decision based on the value of the linear combination of input features.

swarm intelligence Collective behavior of a decentralized, self-organized system, natural or artificial.

synapse The connection (where the weights are) between processing elements in a neural network.

synchronous (real-time) Occurring at the same time.

system architecture The logical and physical design of a system.

system development lifecycle (SDLC) A systematic process for the effective construction of large information systems.

systems dynamics Macro-level simulation models in which aggregate values and trends are considered. The objective is to study the overall behavior of a system over time, rather than the behavior of each individual participant or player in the system.

tacit knowledge Knowledge that is usually in the domain of subjective, cognitive, and experiential learning. It is highly personal and difficult to formalize.

tactical models Models that represent problems for the tactical level (i.e., midlevel) of management.

teleconferencing The use of electronic communication that allows two or more people at different locations to have a simultaneous conference.

TensorFlow A popular open-source deep learning framework originally developed by the Google Brain Group in 2011 as *DistBelief*, and further developed into TensorFlow in 2015.

term–document matrix (TDM) A frequency matrix created from digitized and organized documents (the corpus) where the columns represent the terms while rows represent the individual documents.

text analytics A broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms) as well as information extraction, data mining, and Web mining.

text mining The application of data mining to nonstructured or less structured text files. It entails the generation of meaningful numeric indices from the unstructured text and then processing those indices using various data mining algorithms.

Theano This was developed by the Deep Learning Group at the University of Montreal in 2007 as a Python library to define, optimize, and evaluate mathematical expressions involving multidimensional arrays (i.e., tensors) on CPU or GPU platforms.

theory of certainty factors A theory designed to help incorporate uncertainty into the representation of knowledge (in terms of production rules) for expert systems.

threshold value A hurdle value for the output of a neuron to trigger the next level of neurons. If an output value is smaller than the threshold value, it will not be passed to the next level of neurons.

tokenizing Categorizing a block of text (token) according to the function it performs.

topology The way in which neurons are organized in a neural network.

Torch An open-source scientific computing framework for implementing machine-learning algorithms using GPUs.

tort liability In common law jurisdictions, this is where a wrongful act creates an obligation to pay damages to another.

transformation (transfer) function In a neural network, the function that sums and transforms inputs before a neuron fires. It shows the relationship between the internal activation level and the output of a neuron.

trend analysis The collecting of information and attempting to spot a pattern, or *trend*, in the information.

Turing Test Test to determine whether computers are intelligent when a human interviewer questions a human and a machine and is unable to determine which is which.

uncertainty A decision situation where there is a complete lack of information about what the parameter values are or what the future state of nature will be.

uncontrollable variable (parameter) A factor that affects the result of a decision but is not under the control of the decision maker. These variables can be internal (e.g., related to technology or to policies) or external (e.g., related to legal issues or to climate).

uncontrollable variable A mathematical modeling variable that has to be taken as given—not allowing changes/modifications.

universal basic income (UBI) A proposal to give every citizen a minimum amount of income to ensure no one goes hungry despite the massive loss of jobs that is likely to occur.

unstructured data Data that do not have a predetermined format and are stored in the form of textual documents.

unstructured problem A decision setting where the steps are not entirely fixed or structured, but may require subjective considerations.

unsupervised learning A method of training artificial neural networks in which only input stimuli are shown to the network, which is self-organizing.

user interface The component of a computer system that allows bidirectional communication between the system and its user.

user interface management system (UIMS) The DSS component that handles all interaction between users and the system.

user-developed MSS An MSS developed by one user or by a few users in one department, including decision makers and professionals (i.e., knowledge workers—financial analysts, tax analysts, engineers) who build or use computers to solve problems or enhance their productivity.

utility (on-demand) computing Unlimited computing power and storage capacity that, like electricity, water, and telephone services, can be obtained on demand, used, and reallocated for any application and that are billed on a pay-per-use basis.

vendor-managed inventory (VMI) The practice of retailers making suppliers responsible for determining when to order and how much to order.

video teleconferencing (videoconferencing) Virtual meeting in which participants in one location can see participants at other locations on a large screen or a desktop computer.

virtual (Internet) community A group of people with similar interests who interact with one another using the Internet.

virtual meeting An online meeting whose members are in different locations, possibly in different countries.

virtual personal assistant (VPA) A chatbot that assists individuals by searching for information for them, answering questions, and executing simple tasks. Most well known is Alexa from **Amazon.com**.

virtual team A team whose members are in different places while in a meeting together.

virtual worlds Artificial worlds created by computer systems in which the user has the impression of being immersed.

visual analytics The combination of visualization and predictive analytics.

visual interactive modeling (VIM) A visual model representation technique that allows for user and other system interactions.

visual interactive modeling (VIM) *See* visual interactive simulation (VIS).

visual interactive simulation (VIS) A visual/animated simulation environment that allows for the end user to interact with the model parameters while the mode is running.

visual recognition The addition of some form of computer intelligence and decision making to digitized visual information received from a machine sensor such as a camera.

voice (speech) recognition Translation of human voice into individual words and sentences that are understandable by a computer.

voice of customer (VOC) Applications that focus on "who and how" questions by gathering and reporting direct feedback from site visitors, by benchmarking against other sites and offline channels, and by supporting predictive modeling of future visitor behavior.

voice-over IP (VoIP) Communication systems that transmit voice calls over Internet Protocol (IP)–based networks. Also known as *Internet telephony*.

voice portal A Web site, usually a portal, that has an audio interface.

voice synthesis The technology by which computers convert text to voice (i.e., speak).

Web 2.0 The popular term for advanced Internet technology and applications, including blogs, wikis, RSS, and social bookmarking. One of the most significant differences between Web 2.0 and the traditional World Wide Web is greater collaboration among Internet users and other users, content providers, and enterprises.

Web analytics The application of business analytics activities to Web-based processes, including e-commerce.

Web content mining The extraction of useful information from Web pages.

Web crawlers An application used to read through the content of a Web site automatically.

Web mining The discovery and analysis of interesting and useful information from the Web, about the Web, and usually through Web-based tools.

Web services An architecture that enables assembly of distributed applications from software services and ties them together.

Web structure mining The development of useful information from the links included in Web documents.

Web usage mining The extraction of useful information from the data being generated through Web page visits, transactions, and so on.

Weka A popular, free-of-charge, open-source suite of machine-learning software written in Java, developed at the University of Waikato.

what-if analysis It is an experimental process that helps determine what will happen to the solution/output if an input variable, an assumption, or a parameter value is changed.

wiki A piece of server software available in a Web site that allows users to freely create and edit Web page content using any Web browser.

wikilog A Web log (blog) that allows people to participate as peers; anyone can add, delete, or change content.

word2vec A two-layer neural network that gets a large text corpus as the input and converts each word in the corpus to a numeric vector of any given size, typically ranging from 100 to 1000.

WordNet A popular general-purpose lexicon created at Princeton University.

INDEX

Note: 'A', 'f' and 't' refer to application cases, figures and tables respectively

B

C

D

E

F

G

H

I

J

K

L

M

Q

R

S

T

This page is intentionally left blank

www.pearson.com

ISBN-13: 978-0-13-519201-6
ISBN-10: 0-13-519201-3