CENGAGE

Camm Cochran Fry Ohlmann

Business Analytics

Descriptive • Predictive • Prescriptive

Business Analytics

Descriptive • Predictive • Prescriptive

Jeffrey D. Camm
Wake Forest University

James J. Cochran
University of Alabama

Michael J. Fry
University of Cincinnati

Jeffrey W. Ohlmann
University of Iowa

CENGAGE

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

CENGAGE

***Business Analytics*, Fourth Edition**
Jeffrey D. Camm, James J. Cochran, Michael J. Fry, Jeffrey W. Ohlmann

Senior Vice President, Higher Education & Skills Product: Erin Joyner

Product Director: Jason Fremder

Senior Product Manager: Aaron Arnsparger

Senior Content Manager: Conor Allen

Product Assistant: Maggie Russo

Marketing Manager: Chris Walz

Senior Learning Designer: Brandon Foltz

Digital Delivery Lead: Mark Hopkinson

Intellectual Property Analyst: Ashley Maynard

Intellectual Property Project Manager: Kelli Besse

Production Service: MPS Limited

Senior Project Manager, MPS Limited: Santosh Pandey

Art Director: Chris Doughman

Text Designer: Beckmeyer Design

Cover Designer: Beckmeyer Design

Cover Image: iStockPhoto.com/tawanlubfah

Library of Congress Control Number: 2019921119

ISBN: 978-0-357-13178-7

Loose-leaf Edition:
ISBN: 978-0-357-13179-4

Cengage
200 Pier 4 Boulevard
Boston, MA 02210
USA

Cengage is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com**.

Cengage products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com**.

Printed in the United States of America
Print Number: 01 Print Year: 2020

Brief Contents

ABOUT THE AUTHORS xvii
PREFACE xix

CHAPTER 1	Introduction 1
CHAPTER 2	Descriptive Statistics 19
CHAPTER 3	Data Visualization 85
CHAPTER 4	Probability: An Introduction to Modeling Uncertainty 157
CHAPTER 5	Descriptive Data Mining 213
CHAPTER 6	Statistical Inference 253
CHAPTER 7	Linear Regression 327
CHAPTER 8	Time Series Analysis and Forecasting 407
CHAPTER 9	Predictive Data Mining 459
CHAPTER 10	Spreadsheet Models 509
CHAPTER 11	Monte Carlo Simulation 547
CHAPTER 12	Linear Optimization Models 609
CHAPTER 13	Integer Linear Optimization Models 663
CHAPTER 14	Nonlinear Optimization Models 703
CHAPTER 15	Decision Analysis 737
MULTI-CHAPTER CASE PROBLEMS	
	Capital State University Game-Day Magazines 783
	Hanover Inc. 785
APPENDIX A	Basics of Excel 787
APPENDIX B	Database Basics with Microsoft Access 799
APPENDIX C	Solutions to Even-Numbered Problems (MindTap Reader)

REFERENCES *837*
INDEX *839*

Contents

ABOUT THE AUTHORS xvii
PREFACE xix

CHAPTER 1 Introduction 1

1.1 Decision Making 3
1.2 Business Analytics Defined 4
1.3 A Categorization of Analytical Methods and Models 5
Descriptive Analytics 5
Predictive Analytics 5
Prescriptive Analytics 6
1.4 Big Data 6
Volume 8
Velocity 8
Variety 8
Veracity 8
1.5 Business Analytics in Practice 10
Financial Analytics 10
Human Resource (HR) Analytics 11
Marketing Analytics 11
Health Care Analytics 11
Supply Chain Analytics 12
Analytics for Government and Nonprofits 12
Sports Analytics 12
Web Analytics 13
1.6 Legal and Ethical Issues in the Use of Data and Analytics 13
Summary 16
Glossary 16
Available in the MindTap Reader:
Appendix: Getting Started with R and RStudio
Appendix: Basic Data Manipulation with R

CHAPTER 2 Descriptive Statistics 19

2.1 Overview of Using Data: Definitions and Goals 20
2.2 Types of Data 22
Population and Sample Data 22
Quantitative and Categorical Data 22
Cross-Sectional and Time Series Data 22
Sources of Data 22
2.3 Modifying Data in Excel 25
Sorting and Filtering Data in Excel 25
Conditional Formatting of Data in Excel 28

2.4 Creating Distributions from Data 30
Frequency Distributions for Categorical Data 30
Relative Frequency and Percent Frequency Distributions 31
Frequency Distributions for Quantitative Data 32
Histograms 35
Cumulative Distributions 38
2.5 Measures of Location 40
Mean (Arithmetic Mean) 40
Median 41
Mode 42
Geometric Mean 42
2.6 Measures of Variability 45
Range 45
Variance 46
Standard Deviation 47
Coefficient of Variation 48
2.7 Analyzing Distributions 48
Percentiles 49
Quartiles 50
z-Scores 50
Empirical Rule 51
Identifying Outliers 53
Boxplots 53
2.8 Measures of Association Between Two Variables 56
Scatter Charts 56
Covariance 58
Correlation Coefficient 61
2.9 Data Cleansing 62
Missing Data 62
Blakely Tires 64
Identification of Erroneous Outliers and Other Erroneous Values 66
Variable Representation 68
Summary 69
Glossary 70
Problems 71
Case Problem 1: Heavenly Chocolates Web Site Transactions 81
Case Problem 2: African Elephant Populations 82
Available in the MindTap Reader:
Appendix: Descriptive Statistics with R

CHAPTER 3 Data Visualization 85
3.1 Overview of Data Visualization 88
Effective Design Techniques 88
3.2 Tables 91
Table Design Principles 92
Crosstabulation 93

PivotTables in Excel 96
Recommended PivotTables in Excel 100
3.3 Charts 102
Scatter Charts 102
Recommended Charts in Excel 104
Line Charts 105
Bar Charts and Column Charts 109
A Note on Pie Charts and Three-Dimensional Charts 110
Bubble Charts 112
Heat Maps 113
Additional Charts for Multiple Variables 115
PivotCharts in Excel 118
3.4 Advanced Data Visualization 120
Advanced Charts 120
Geographic Information Systems Charts 123
3.5 Data Dashboards 125
Principles of Effective Data Dashboards 125
Applications of Data Dashboards 126
Summary 128
Glossary 128
Problems 129
Case Problem 1: Pelican stores 139
Case Problem 2: Movie Theater Releases 140
Appendix: Data Visualization in Tableau 141
Available in the MindTap Reader:
Appendix: Creating Tabular and Graphical Presentations with R

CHAPTER 4 Probability: An Introduction to Modeling Uncertainty 157
4.1 Events and Probabilities 159
4.2 Some Basic Relationships of Probability 160
Complement of an Event 160
Addition Law 161
4.3 Conditional Probability 163
Independent Events 168
Multiplication Law 168
Bayes' Theorem 169
4.4 Random Variables 171
Discrete Random Variables 171
Continuous Random Variables 172
4.5 Discrete Probability Distributions 173
Custom Discrete Probability Distribution 173
Expected Value and Variance 175
Discrete Uniform Probability Distribution 178
Binomial Probability Distribution 179
Poisson Probability Distribution 182

4.6 Continuous Probability Distributions 185
Uniform Probability Distribution 185
Triangular Probability Distribution 187
Normal Probability Distribution 189
Exponential Probability Distribution 194
Summary 198
Glossary 198
Problems 200
Case Problem 1: Hamilton County Judges 209
Case Problem 2: McNeil's Auto Mall 210
Case Problem 3: Gebhardt Electronics 211
Available in the MindTap Reader:
Appendix: Discrete Probability Distributions with R
Appendix: Continuous Probability Distributions with R

CHAPTER 5 **Descriptive Data Mining 213**
5.1 Cluster Analysis 215
Measuring Distance Between Observations 215
k-Means Clustering 218
Hierarchical Clustering and Measuring Dissimilarity Between Clusters 221
Hierarchical Clustering Versus *k*-Means Clustering 225
5.2 Association Rules 226
Evaluating Association Rules 228
5.3 Text Mining 229
Voice of the Customer at Triad Airline 229
Preprocessing Text Data for Analysis 231
Movie Reviews 232
Computing Dissimilarity Between Documents 234
Word Clouds 234
Summary 235
Glossary 235
Problems 237
Case Problem 1: Big Ten Expansion 251
Case Problem 2: Know Thy Customer 251
Available in the MindTap Reader:
Appendix: Getting Started with Rattle in R
Appendix: *k*-Means Clustering with R
Appendix: Hierarchical Clustering with R
Appendix: Association Rules with R
Appendix: Text Mining with R
Appendix: R/Rattle Settings to Solve Chapter 5 Problems
Appendix: Opening and Saving Excel Files in JMP Pro
Appendix: Hierarchical Clustering with JMP Pro

Appendix: *k*-Means Clustering with JMP Pro
Appendix: Association Rules with JMP Pro
Appendix: Text Mining with JMP Pro
Appendix: JMP Pro Settings to Solve Chapter 5 Problems

CHAPTER 6 **Statistical Inference 253**

6.1 Selecting a Sample 256
- Sampling from a Finite Population 256
- Sampling from an Infinite Population 257

6.2 Point Estimation 260
- Practical Advice 262

6.3 Sampling Distributions 262
- Sampling Distribution of $\bar{x}$ 265
- Sampling Distribution of $\bar{p}$ 270

6.4 Interval Estimation 273
- Interval Estimation of the Population Mean 273
- Interval Estimation of the Population Proportion 280

6.5 Hypothesis Tests 283
- Developing Null and Alternative Hypotheses 283
- Type I and Type II Errors 286
- Hypothesis Test of the Population Mean 287
- Hypothesis Test of the Population Proportion 298

6.6 Big Data, Statistical Inference, and Practical Significance 301
- Sampling Error 301
- Nonsampling Error 302
- Big Data 303
- Understanding What Big Data Is 304
- Big Data and Sampling Error 305
- Big Data and the Precision of Confidence Intervals 306
- Implications of Big Data for Confidence Intervals 307
- Big Data, Hypothesis Testing, and *p* Values 308
- Implications of Big Data in Hypothesis Testing 310

Summary 310
Glossary 311
Problems 314
Case Problem 1: Young Professional Magazine 324
Case Problem 2: Quality Associates, Inc. 325

Available in the MindTap Reader:
Appendix: Random Sampling with R
Appendix: Interval Estimation with R
Appendix: Hypothesis Testing with R

CHAPTER 7 **Linear Regression 327**

7.1 Simple Linear Regression Model 329
- Regression Model 329
- Estimated Regression Equation 329

7.2 Least Squares Method 331
Least Squares Estimates of the Regression Parameters 333
Using Excel's Chart Tools to Compute the Estimated Regression Equation 335
7.3 Assessing the Fit of the Simple Linear Regression Model 337
The Sums of Squares 337
The Coefficient of Determination 339
Using Excel's Chart Tools to Compute the Coefficient of Determination 340
7.4 The Multiple Regression Model 341
Regression Model 341
Estimated Multiple Regression Equation 341
Least Squares Method and Multiple Regression 342
Butler Trucking Company and Multiple Regression 342
Using Excel's Regression Tool to Develop the Estimated Multiple Regression Equation 343
7.5 Inference and Regression 346
Conditions Necessary for Valid Inference in the Least Squares Regression Model 347
Testing Individual Regression Parameters 351
Addressing Nonsignificant Independent Variables 354
Multicollinearity 355
7.6 Categorical Independent Variables 358
Butler Trucking Company and Rush Hour 358
Interpreting the Parameters 360
More Complex Categorical Variables 361
7.7 Modeling Nonlinear Relationships 363
Quadratic Regression Models 364
Piecewise Linear Regression Models 368
Interaction Between Independent Variables 370
7.8 Model Fitting 375
Variable Selection Procedures 375
Overfitting 376
7.9 Big Data and Regression 377
Inference and Very Large Samples 377
Model Selection 380
7.10 Prediction with Regression 382
Summary 384
Glossary 384
Problems 386
Case Problem 1: Alumni Giving 402
Case Problem 2: Consumer Research, Inc. 404
Case Problem 3: Predicting Winnings for NASCAR Drivers 405
Available in the MindTap Reader:
Appendix: Simple Linear Regression with R

Appendix: Multiple Linear Regression with R
Appendix: Regression Variable Selection Procedures with R

CHAPTER 8 **Time Series Analysis and Forecasting 407**

8.1 Time Series Patterns 410
Horizontal Pattern 410
Trend Pattern 412
Seasonal Pattern 413
Trend and Seasonal Pattern 414
Cyclical Pattern 417
Identifying Time Series Patterns 417
8.2 Forecast Accuracy 417
8.3 Moving Averages and Exponential Smoothing 421
Moving Averages 422
Exponential Smoothing 426
8.4 Using Regression Analysis for Forecasting 430
Linear Trend Projection 430
Seasonality Without Trend 432
Seasonality with Trend 433
Using Regression Analysis as a Causal Forecasting Method 436
Combining Causal Variables with Trend and Seasonality Effects 439
Considerations in Using Regression in Forecasting 440
8.5 Determining the Best Forecasting Model to Use 440
Summary 441
Glossary 441
Problems 442
Case Problem 1: Forecasting Food and Beverage Sales 450
Case Problem 2: Forecasting Lost Sales 450
Appendix: Using the Excel Forecast Sheet 452
Available in the MindTap Reader:
Appendix: Forecasting with R

CHAPTER 9 **Predictive Data Mining 459**

9.1 Data Sampling, Preparation, and Partitioning 461
Static Holdout Method 461
k-Fold Cross-Validation 462
Class Imbalanced Data 463
9.2 Performance Measures 464
Evaluating the Classification of Categorical Outcomes 464
Evaluating the Estimation of Continuous Outcomes 470
9.3 Logistic Regression 471
9.4 *k*-Nearest Neighbors 475
Classifying Categorical Outcomes with *k*-Nearest Neighbors 475
Estimating Continuous Outcomes with *k*-Nearest Neighbors 477

9.5 Classification and Regression Trees 478
Classifying Categorical Outcomes with a Classification Tree 478
Estimating Continuous Outcomes with a Regression Tree 483
Ensemble Methods 485
Summary 489
Glossary 491
Problems 492
Case Problem: Grey Code Corporation 505
Available in the MindTap Reader:
Appendix: Classification via Logistic Regression with R
Appendix: *k*-Nearest Neighbor Classification with R
Appendix: *k*-Nearest Neighbor Regression with R
Appendix: Individual Classification Trees with R
Appendix: Individual Regression Trees with R
Appendix: Random Forests of Classification Trees with R
Appendix: Random Forests of Regression Trees with R
Appendix: R/Rattle Settings to Solve Chapter 9 Problems
Appendix: Data Partitioning with JMP Pro
Appendix: Classification via Logistic Regression with JMP Pro
Appendix: *k*-Nearest Neighbors Classification and Regression with JMP Pro
Appendix: Individual Classification and Regression Trees with JMP Pro
Appendix: Random Forests of Classification or Regression Trees with JMP Pro
Appendix: JMP Pro Settings to Solve Chapter 9 Problems

CHAPTER 10 **Spreadsheet Models 509**
10.1 Building Good Spreadsheet Models 511
Influence Diagrams 511
Building a Mathematical Model 511
Spreadsheet Design and Implementing the Model in a Spreadsheet 513
10.2 What-If Analysis 516
Data Tables 516
Goal Seek 518
Scenario Manager 520
10.3 Some Useful Excel Functions for Modeling 525
SUM and SUMPRODUCT 526
IF and COUNTIF 528
VLOOKUP 530
10.4 Auditing Spreadsheet Models 532
Trace Precedents and Dependents 532
Show Formulas 532
Evaluate Formulas 534
Error Checking 534
Watch Window 535

10.5 Predictive and Prescriptive Spreadsheet Models 536
Summary 537
Glossary 537
Problems 538
Case Problem: Retirement Plan 544

CHAPTER 11 **Monte Carlo Simulation 547**
11.1 Risk Analysis for Sanotronics LLC 549
Base-Case Scenario 549
Worst-Case Scenario 550
Best-Case Scenario 550
Sanotronics Spreadsheet Model 550
Use of Probability Distributions to Represent Random Variables 551
Generating Values for Random Variables with Excel 553
Executing Simulation Trials with Excel 557
Measuring and Analyzing Simulation Output 557
11.2 Inventory Policy Analysis for Promus Corp 561
Spreadsheet Model for Promus 562
Generating Values for Promus Corp's Demand 563
Executing Simulation Trials and Analyzing Output 565
11.3 Simulation Modeling for Land Shark Inc. 568
Spreadsheet Model for Land Shark 569
Generating Values for Land Shark's Random Variables 570
Executing Simulation Trials and Analyzing Output 572
Generating Bid Amounts with Fitted Distributions 575
11.4 Simulation with Dependent Random Variables 580
Spreadsheet Model for Press Teag Worldwide 580
11.5 Simulation Considerations 585
Verification and Validation 585
Advantages and Disadvantages of Using Simulation 585
Summary 586
Summary of Steps for Conducting a Simulation Analysis 586
Glossary 587
Problems 587
Case Problem: Four Corners 600
Appendix: Common Probability Distributions for Simulation 602

CHAPTER 12 **Linear Optimization Models 609**
12.1 A Simple Maximization Problem 611
Problem Formulation 612
Mathematical Model for the Par, Inc. Problem 614
12.2 Solving the Par, Inc. Problem 614
The Geometry of the Par, Inc. Problem 615
Solving Linear Programs with Excel Solver 617

12.3 A Simple Minimization Problem 621
Problem Formulation 621
Solution for the M&D Chemicals Problem 621
12.4 Special Cases of Linear Program Outcomes 623
Alternative Optimal Solutions 624
Infeasibility 625
Unbounded 626
12.5 Sensitivity Analysis 628
Interpreting Excel Solver Sensitivity Report 628
12.6 General Linear Programming Notation and More Examples 630
Investment Portfolio Selection 631
Transportation Planning 633
Maximizing Banner Ad Revenue 637
12.7 Generating an Alternative Optimal Solution for a Linear Program 642
Summary 644
Glossary 645
Problems 646
Case Problem: Investment Strategy 660

CHAPTER 13 Integer Linear Optimization Models 663
13.1 Types of Integer Linear Optimization Models 664
13.2 Eastborne Realty, an Example of Integer Optimization 665
The Geometry of Linear All-Integer Optimization 666
13.3 Solving Integer Optimization Problems with Excel Solver 668
A Cautionary Note About Sensitivity Analysis 671
13.4 Applications Involving Binary Variables 673
Capital Budgeting 673
Fixed Cost 675
Bank Location 678
Product Design and Market Share Optimization 680
13.5 Modeling Flexibility Provided by Binary Variables 683
Multiple-Choice and Mutually Exclusive Constraints 683
k Out of *n* Alternatives Constraint 684
Conditional and Corequisite Constraints 684
13.6 Generating Alternatives in Binary Optimization 685
Summary 687
Glossary 688
Problems 689
Case Problem: Applecore Children's Clothing 701

CHAPTER 14 Nonlinear Optimization Models 703
14.1 A Production Application: Par, Inc. Revisited 704
An Unconstrained Problem 704
A Constrained Problem 705
Solving Nonlinear Optimization Models Using Excel Solver 707
Sensitivity Analysis and Shadow Prices in Nonlinear Models 708

14.2 Local and Global Optima 709
Overcoming Local Optima with Excel Solver 712
14.3 A Location Problem 714
14.4 Markowitz Portfolio Model 715
14.5 Adoption of a New Product: The Bass Forecasting Model 720
Summary 723
Glossary 724
Problems 724
Case Problem: Portfolio Optimization with Transaction Costs 732

CHAPTER 15 **Decision Analysis 737**
15.1 Problem Formulation 739
Payoff Tables 740
Decision Trees 740
15.2 Decision Analysis Without Probabilities 741
Optimistic Approach 741
Conservative Approach 742
Minimax Regret Approach 742
15.3 Decision Analysis with Probabilities 744
Expected Value Approach 744
Risk Analysis 746
Sensitivity Analysis 747
15.4 Decision Analysis with Sample Information 748
Expected Value of Sample Information 753
Expected Value of Perfect Information 753
15.5 Computing Branch Probabilities with Bayes' Theorem 754
15.6 Utility Theory 757
Utility and Decision Analysis 758
Utility Functions 762
Exponential Utility Function 765
Summary 767
Glossary 767
Problems 769
Case Problem: Property Purchase Strategy 780

MULTI-CHAPTER CASE PROBLEMS
Capital State University Game-Day Magazines 783
Hanover Inc. 785

APPENDIX A Basics of Excel 787

APPENDIX B Database Basics with Microsoft Access 799

APPENDIX C Solutions to Even-Numbered Problems (MindTap Reader)

REFERENCES 837

INDEX 839

About the Authors

Jeffrey D. Camm. is the Inmar Presidential Chair and Associate Dean of Business Analytics in the School of Business at Wake Forest University. Born in Cincinnati, Ohio, he holds a B.S. from Xavier University (Ohio) and a Ph.D. from Clemson University. Prior to joining the faculty at Wake Forest, he was on the faculty of the University of Cincinnati. He has also been a visiting scholar at Stanford University and a visiting professor of business administration at the Tuck School of Business at Dartmouth College.

Dr. Camm has published over 40 papers in the general area of optimization applied to problems in operations management and marketing. He has published his research in *Science*, *Management Science*, *Operations Research*, *Interfaces*, and other professional journals. Dr. Camm was named the Dornoff Fellow of Teaching Excellence at the University of Cincinnati and he was the 2006 recipient of the INFORMS Prize for the Teaching of Operations Research Practice. A firm believer in practicing what he preaches, he has served as an operations research consultant to numerous companies and government agencies. From 2005 to 2010 he served as editor-in-chief of *Interfaces*. In 2016, Professor Camm received the George E. Kimball Medal for service to the operations research profession, and in 2017 he was named an INFORMS Fellow.

James J. Cochran. James J. Cochran is Associate Dean for Research, Professor of Applied Statistics and the Rogers-Spivey Faculty Fellow at The University of Alabama. Born in Dayton, Ohio, he earned his B.S., M.S., and M.B.A. from Wright State University and his Ph.D. from the University of Cincinnati. He has been at The University of Alabama since 2014 and has been a visiting scholar at Stanford University, Universidad de Talca, the University of South Africa and Pole Universitaire Leonard de Vinci.

Dr. Cochran has published more than 40 papers in the development and application of operations research and statistical methods. He has published in several journals, including *Management Science*, *The American Statistician*, *Communications in Statistics—Theory and Methods*, *Annals of Operations Research*, *European Journal of Operational Research*, *Journal of Combinatorial Optimization*, *Interfaces*, and *Statistics and Probability Letters*. He received the 2008 INFORMS Prize for the Teaching of Operations Research Practice, 2010 Mu Sigma Rho Statistical Education Award and 2016 Waller Distinguished Teaching Career Award from the American Statistical Association. Dr. Cochran was elected to the International Statistics Institute in 2005, named a Fellow of the American Statistical Association in 2011, and named a Fellow of INFORMS in 2017. He also received the Founders Award in 2014 and the Karl E. Peace Award in 2015 from the American Statistical Association, and he received the INFORMS President's Award in 2019.

A strong advocate for effective operations research and statistics education as a means of improving the quality of applications to real problems, Dr. Cochran has chaired teaching effectiveness workshops around the globe. He has served as an operations research consultant to numerous companies and not-for-profit organizations. He served as editor-in-chief of *INFORMS Transactions on Education* and is on the editorial board of *INFORMS Journal of Applied Analytics*, *International Transactions in Operational Research*, and *Significance*.

Michael J. Fry. Michael J. Fry is Professor of Operations, Business Analytics, and Information Systems (OBAIS) and Academic Director of the Center for Business Analytics in the Carl H. Lindner College of Business at the University of Cincinnati. Born in Killeen, Texas, he earned a B.S. from Texas A&M University, and M.S.E. and Ph.D. degrees from the University of Michigan. He has been at the University of Cincinnati since 2002, where he served as Department Head from 2014 to 2018 and has been named a Lindner Research Fellow. He has also been a visiting professor at Cornell University and at the University of British Columbia.

Professor Fry has published more than 25 research papers in journals such as *Operations Research*, *Manufacturing & Service Operations Management*, *Transportation Science*, *Naval Research Logistics*, *IIE Transactions*, *Critical Care Medicine*, and *Interfaces*. He serves on editorial boards for journals such as *Production and Operations Management*, *INFORMS Journal of Applied Analytics* (formerly *Interfaces*), and *Journal of Quantitative Analysis in Sports*. His research interests are in applying analytics to the areas of supply chain management, sports, and public-policy operations. He has worked with many different organizations for his research, including Dell, Inc., Starbucks Coffee Company, Great American Insurance Group, the Cincinnati Fire Department, the State of Ohio Election Commission, the Cincinnati Bengals, and the Cincinnati Zoo & Botanical Gardens. In 2008, he was named a finalist for the Daniel H. Wagner Prize for Excellence in Operations Research Practice, and he has been recognized for both his research and teaching excellence at the University of Cincinnati. In 2019, he led the team that was awarded the INFORMS UPS George D. Smith Prize on behalf of the OBAIS Department at the University of Cincinnati.

Jeffrey W. Ohlmann. Jeffrey W. Ohlmann is Associate Professor of Business Analytics and Huneke Research Fellow in the Tippie College of Business at the University of Iowa. Born in Valentine, Nebraska, he earned a B.S. from the University of Nebraska, and M.S. and Ph.D. degrees from the University of Michigan. He has been at the University of Iowa since 2003.

Professor Ohlmann's research on the modeling and solution of decision-making problems has produced more than two dozen research papers in journals such as *Operations Research*, *Mathematics of Operations Research*, *INFORMS Journal on Computing*, *Transportation Science*, and the *European Journal of Operational Research*. He has collaborated with companies such as Transfreight, LeanCor, Cargill, the Hamilton County Board of Elections, and three National Football League franchises. Because of the relevance of his work to industry, he was bestowed the George B. Dantzig Dissertation Award and was recognized as a finalist for the Daniel H. Wagner Prize for Excellence in Operations Research Practice.

Preface

Business Analytics 4E is designed to introduce the concept of business analytics to undergraduate and graduate students. This edition builds upon what was one of the first collections of materials that are essential to the growing field of business analytics. In Chapter 1, we present an overview of business analytics and our approach to the material in this textbook. In simple terms, business analytics helps business professionals make better decisions based on data. We discuss models for summarizing, visualizing, and understanding useful information from historical data in Chapters 2 through 6. Chapters 7 through 9 introduce methods for both gaining insights from historical data and predicting possible future outcomes. Chapter 10 covers the use of spreadsheets for examining data and building decision models. In Chapter 11, we demonstrate how to explicitly introduce uncertainty into spreadsheet models through the use of Monte Carlo simulation. In Chapters 12 through 14, we discuss optimization models to help decision makers choose the best decision based on the available data. Chapter 15 is an overview of decision analysis approaches for incorporating a decision maker's views about risk into decision making. In Appendix A we present optional material for students who need to learn the basics of using Microsoft Excel. The use of databases and manipulating data in Microsoft Access is discussed in Appendix B. Appendixes in many chapters illustrate the use of additional software tools such as R, JMP Pro and Tableau to apply analytics methods.

This textbook can be used by students who have previously taken a course on basic statistical methods as well as students who have not had a prior course in statistics. *Business Analytics* 4E is also amenable to a two-course sequence in business statistics and analytics. All statistical concepts contained in this textbook are presented from a business analytics perspective using practical business examples. Chapters 2, 4, 6, and 7 provide an introduction to basic statistical concepts that form the foundation for more advanced analytics methods. Chapters 3, 5, and 9 cover additional topics of data visualization and data mining that are not traditionally part of most introductory business statistics courses, but they are exceedingly important and commonly used in current business environments. Chapter 10 and Appendix A provide the foundational knowledge students need to use Microsoft Excel for analytics applications. Chapters 11 through 15 build upon this spreadsheet knowledge to present additional topics that are used by many organizations that are leaders in the use of prescriptive analytics to improve decision making.

Updates in the Fourth Edition

The fourth edition of *Business Analytics* is a major revision. We have added online appendixes for many topics in Chapters 1 through 9 that introduce the use of R, the exceptionally popular open-source software for analytics. *Business Analytics* 4E also includes an appendix to Chapter 3 introducing the powerful data visualization software Tableau. We have further enhanced our data mining chapters to allow instructors to choose their preferred means of teaching this material in terms of software usage. We have expanded the number of conceptual homework problems in both Chapters 5 and 9 to increase the number of opportunities for students learn about data mining and solve problems without the use of data mining software. Additionally, we now include online appendixes on using JMP Pro and R for teaching data mining so that instructors can choose their favored way of teaching this material. Other changes in this edition include an expanded discussion of binary variables for integer optimization in Chapter 13, an additional example in Chapter 11 for Monte Carlo simulation, and new and revised homework problems and cases.

- **Tableau Appendix for Data Visualization.** Chapter 3 now includes a new appendix that introduces the use of the software Tableau for data visualization. Tableau is a very powerful software for creating meaningful data visualizations that can be used to display, and to analyze, data. The appendix includes step-by-step directions for generating many of the charts used in Chapters 2 and 3 in Tableau.

- **Incorporation of R.** R is an exceptionally powerful open-source software that is widely used for a variety of statistical and analytics methods. We now include online appendixes that introduce the use of R for many of the topics covered in Chapters 1 through 9, including data visualization and data mining. These appendixes include step-by-step directions for using R to implement the methods described in these chapters. To facilitate the use of R, we introduce RStudio, an open-source integrated development environment (IDE) that provides a menu-driven interface for R. For Chapters 5 and 9 that cover data mining, we introduce the use of Rattle, a library package providing a graphical-user interface for R specifically tailored for data mining functionality. The use of RStudio and Rattle eases the learning curve of using R so that students can focus on learning the methods and interpreting the output.
- **Updates for Data Mining Chapters.** Chapters 5 and 9 have received extensive updates. We have moved the Descriptive Data Mining chapter to Chapter 5 so that it is located after our chapter on Probability. This allows us to use probability concepts such as conditional probability to explain association rule measures. Additional content on text mining and further discussion of ways to measure distance between observations have been added to a reorganized Descriptive Data Mining chapter. Descriptions of cross-validation approaches, methods of addressing class imbalanced data, and out-of-bag estimation in ensemble methods have been added to Chapter 9 on Predictive Data Mining. The end-of-chapter problems in Chapters 5 and 9 have been revised and generalized to accommodate the use of a wide range of data mining software. To allow instructors to choose different software for use with these chapters, we have created online appendixes for both JMP Pro and R. JMP has introduced a new version of its software (JMP Pro 14) since the previous edition of this textbook, so we have updated our JMP Pro output and step-by-step instructions to reflect changes in this software. We have also written online appendixes for Chapters 5 and 9 that use R and the graphical-user interface Rattle to introduce topics from these chapters to students. The use of Rattle removes some of the more difficult line-by-line coding in R to perform many common data mining techniques so that students can concentrate on learning the methods rather than coding syntax. For some data mining techniques that are not available in Rattle, we show how to accomplish these methods using R code. And for all of our textbook examples, we include the exact R code that can be used to solve the examples. We have also added homework problems to Chapters 5 and 9 that can be solved without using any specialized software. This allows instructors to cover the basics of data mining without introducing any additional software. The online appendixes for Chapters 5 and 9 also include JMP Pro and R specific instructions for how to solve the end-of-chapter problems and cases using JMP Pro and R. Problem and case solutions using both JMP Pro and R are also available to instructors.
- **Additional Simulation Model Example.** We have added an additional example of a simulation model in Chapter 11. This new example helps bridge the gap in the difficultly levels of the previous examples. The new example also gives students additional information on how to build and interpret simulation models.
- **New Cases.** *Business Analytics* 4E includes nine new end-of-chapter cases that allow students to work on more extensive problems related to the chapter material and work with larger data sets. We have also written two new cases that require the use of material from multiple chapters. This helps students understand the connections between the material in different chapters and is more representative of analytics projects in practice where the methods used are often not limited to a single type.
- **Legal and Ethical Issues Related to Analytics and Big Data.** Chapter 1 now includes a section that discusses legal and ethical issues related to analytics and the use of big data. This section discusses legal issues related to the protection of data as well as ethical issues related to the misuse and unintended consequences of analytics applications.

- **New End-of-Chapter Problems.** The fourth edition of this textbook includes more than 20 new problems. We have also revised many of the existing problems to update and improve clarity. Each end-of-chapter problem now also includes a short header to make the application of the exercise more clear. As we have done in past editions, Excel solution files are available to instructors for problems that require the use of Excel. For problems that require the use of software in the data-mining chapters (Chapters 5 and 9), we include solutions for both JMP Pro and R/Rattle.

Continued Features and Pedagogy

In the fourth edition of this textbook, we continue to offer all of the features that have been successful in the first two editions. Some of the specific features that we use in this textbook are listed below.

- **Integration of Microsoft Excel:** Excel has been thoroughly integrated throughout this textbook. For many methodologies, we provide instructions for how to perform calculations both by hand and with Excel. In other cases where realistic models are practical only with the use of a spreadsheet, we focus on the use of Excel to describe the methods to be used.
- **Notes and Comments:** At the end of many sections, we provide Notes and Comments to give the student additional insights about the methods presented in that section. These insights include comments on the limitations of the presented methods, recommendations for applications, and other matters. Additionally, margin notes are used throughout the textbook to provide additional insights and tips related to the specific material being discussed.
- **Analytics in Action:** Each chapter contains an Analytics in Action article. Several of these have been updated and replaced for the fourth edition. These articles present interesting examples of the use of business analytics in practice. The examples are drawn from many different organizations in a variety of areas including healthcare, finance, manufacturing, marketing, and others.
- **DATAfiles and MODELfiles:** All data sets used as examples and in student exercises are also provided online on the companion site as files available for download by the student. DATAfiles are Excel files (or .csv files for easy import into JMP Pro and R/Rattle) that contain data needed for the examples and problems given in the textbook. MODELfiles contain additional modeling features such as extensive use of Excel formulas or the use of Excel Solver, JMP Pro, or R.
- **Problems and Cases:** With the exception of Chapter 1, each chapter contains an extensive selection of problems to help the student master the material presented in that chapter. The problems vary in difficulty and most relate to specific examples of the use of business analytics in practice. Answers to even-numbered problems are provided in an online supplement for student access. With the exception of Chapter 1, each chapter also includes at least one in-depth case study that connects many of the different methods introduced in the chapter. The case studies are designed to be more open-ended than the chapter problems, but enough detail is provided to give the student some direction in solving the cases. New to the fourth edition is the inclusion of two cases that require the use of material from multiple chapters in the text to better illustrate how concepts from different chapters relate to each other.

MindTap

MindTap is a customizable digital course solution that includes an interactive eBook, autograded exercises from the textbook, algorithmic practice problems with solutions feedback, Exploring Analytics visualizations, Adaptive Test Prep, and more! MindTap is also where instructors and users can find the online appendixes for JMP Pro and R/Rattle. All of these materials offer students better access to resources to understand the materials within the course. For more information on MindTap, please contact your Cengage representative.

WebAssign

Prepare for class with confidence using WebAssign from Cengage. This online learning platform fuels practice, so students can truly absorb what you learn – and are better prepared come test time. Videos, Problem Walk-Throughs, and End-of-Chapter problems and cases with instant feedback help them understand the important concepts, while instant grading allows you and them to see where they stand in class. Class Insights allows students to see what topics they have mastered and which they are struggling with, helping them identify where to spend extra time. Study Smarter with WebAssign.

For Students

Online resources are available to help the student work more efficiently. The resources can be accessed through **www.cengage.com/decisionsciences/camm/ba/4e**.

- **R, RStudio, and Rattle:** R, RStudio, and Rattle are open-source software, so they are free to download. *Business Analytics* 4E includes step-by-step instructions for downloading these software.
- **JMP Pro:** Many universities have site licenses of SAS Institute's JMP Pro software on both Mac and Windows. These are typically offered through your university's software licensing administrator. Faculty may contact the JMP Academic team to find out if their universities have a license or to request a complementary instructor copy at www.jmp.com/contact-academic. For institutions without a site license, students may rent a 6- or 12-month license for JMP at www.onthehub.com/jmp.
- **Data Files:** A complete download of all data files associated with this text.

For Instructors

Instructor resources are available to adopters on the Instructor Companion Site, which can be found and accessed at **www.cengage.com/decisionsciences/camm/ba/4e** including:

- **Solutions Manual:** The Solutions Manual, prepared by the authors, includes solutions for all problems in the text. It is available online as well as print. Excel solution files are available to instructors for those problems that require the use of Excel. Solutions for Chapters 5 and 9 are available using both JMP Pro and R/Rattle for data mining problems.
- **Solutions to Case Problems:** These are also prepared by the authors and contain solutions to all case problems presented in the text. Case solutions for Chapters 5 and 9 are provided using both JMP Pro and R/Rattle. Extensive case solutions are also provided for the new multi-chapter cases that draw on material from multiple chapters.
- **PowerPoint Presentation Slides:** The presentation slides contain a teaching outline that incorporates figures to complement instructor lectures.
- **Test Bank:** Cengage Learning Testing Powered by Cognero is a flexible, online system that allows you to:
    - author, edit, and manage test bank content from multiple Cengage Learning solutions,
    - create multiple test versions in an instant, and
    - deliver tests from your Learning Management System (LMS), your classroom, or wherever you want.

Acknowledgments

We would like to acknowledge the work of reviewers and users who have provided comments and suggestions for improvement of this text. Thanks to:

Rafael Becerril Arreola
University of South Carolina

Matthew D. Bailey
Bucknell University

Phillip Beaver
University of Denver

M. Khurrum S. Bhutta
Ohio University

Paolo Catasti
Virginia Commonwealth University

Q B. Chung
Villanova University

Elizabeth A. Denny
University of Kentucky

Mike Taein Eom
University of Portland

Yvette Njan Essounga
Fayetteville State University

Lawrence V. Fulton
Texas State University

Tom Groleau
Carthage College

James F. Hoelscher
Lincoln Memorial University

Eric Huggins
Fort Lewis College

Faizul Huq
Ohio University

Marco Lam
York College of Pennsylvania

Thomas Lee
University of California, Berkeley

Roger Myerson
Northwestern University

Ram Pakath
University of Kentucky

Susan Palocsay
James Madison University

Andy Shogan
University of California, Berkeley

Dothan Truong
Embry-Riddle Aeronautical University

Kai Wang
Wake Technical Community College

Ed Wasil
American University

Ed Winkofsky
University of Cincinnati

A special thanks goes to our associates from business and industry who supplied the Analytics in Action features. We recognize them individually by a credit line in each of the articles. We are also indebted to our senior product manager, Aaron Arnsparger; our Senior Content Manager, Conor Allen; senior learning designer, Brandon Foltz; digital delivery lead, Mark Hopkinson; and our senior project manager at MPS Limited, Santosh Pandey, for their editorial counsel and support during the preparation of this text.

Jeffrey D. Camm
James J. Cochran
Michael J. Fry
Jeffrey W. Ohlmann

Chapter 1

Introduction

CONTENTS

1.1 DECISION MAKING

1.2 BUSINESS ANALYTICS DEFINED

1.3 A CATEGORIZATION OF ANALYTICAL METHODS AND MODELS
- Descriptive Analytics
- Predictive Analytics
- Prescriptive Analytics

1.4 BIG DATA
- Volume
- Velocity
- Variety
- Veracity

1.5 BUSINESS ANALYTICS IN PRACTICE
- Financial Analytics
- Human Resource (HR) Analytics
- Marketing Analytics
- Health Care Analytics
- Supply Chain Analytics
- Analytics for Government and Nonprofits
- Sports Analytics
- Web Analytics

1.6 LEGAL AND ETHICAL ISSUES IN THE USE OF DATA AND ANALYTICS

SUMMARY 16
GLOSSARY 16

AVAILABLE IN THE MINDTAP READER:
- APPENDIX: GETTING STARTED WITH R AND RSTUDIO
- APPENDIX: BASIC DATA MANIPULATION WITH R

You apply for a loan for the first time. How does the bank assess the riskiness of the loan it might make to you? How does Amazon.com know which books and other products to recommend to you when you log in to their web site? How do airlines determine what price to quote to you when you are shopping for a plane ticket? How can doctors better diagnose and treat you when you are ill or injured?

You may be applying for a loan for the first time, but millions of people around the world have applied for loans before. Many of these loan recipients have paid back their loans in full and on time, but some have not. The bank wants to know whether you are more like those who have paid back their loans or more like those who defaulted. By comparing your credit history, financial situation, and other factors to the vast database of previous loan recipients, the bank can effectively assess how likely you are to default on a loan.

Similarly, Amazon.com has access to data on millions of purchases made by customers on its web site. Amazon.com examines your previous purchases, the products you have viewed, and any product recommendations you have provided. Amazon.com then searches through its huge database for customers who are similar to you in terms of product purchases, recommendations, and interests. Once similar customers have been identified, their purchases form the basis of the recommendations given to you.

Prices for airline tickets are frequently updated. The price quoted to you for a flight between New York and San Francisco today could be very different from the price that will be quoted tomorrow. These changes happen because airlines use a pricing strategy known as revenue management. Revenue management works by examining vast amounts of data on past airline customer purchases and using these data to forecast future purchases. These forecasts are then fed into sophisticated optimization algorithms that determine the optimal price to charge for a particular flight and when to change that price. Revenue management has resulted in substantial increases in airline revenues.

Finally, consider the case of being evaluated by a doctor for a potentially serious medical issue. Hundreds of medical papers may describe research studies done on patients facing similar diagnoses, and thousands of data points exist on their outcomes. However, it is extremely unlikely that your doctor has read every one of these research papers or is aware of all previous patient outcomes. Instead of relying only on her medical training and knowledge gained from her limited set of previous patients, wouldn't it be better for your doctor to have access to the expertise and patient histories of thousands of doctors around the world?

A group of IBM computer scientists initiated a project to develop a new decision technology to help in answering these types of questions. That technology is called Watson, named after the founder of IBM, Thomas J. Watson. The team at IBM focused on one aim: How the vast amounts of data now available on the Internet can be used to make more data-driven, smarter decisions. Watson is an example of the exploding field of **artificial intelligence (AI)**. Broadly speaking, AI is the use of data and computers to make decisions that would have in the past required human intelligence. Often, the computer software mimics the way we understand the human brain functions.

Watson became a household name in 2011, when it famously won the television game show, *Jeopardy!* Since that proof of concept in 2011, IBM has reached agreements with the health insurance provider WellPoint (now part of Anthem), the financial services company Citibank, Memorial Sloan-Kettering Cancer Center, and automobile manufacturer General Motors to apply Watson to the decision problems that they face.

Watson is a system of computing hardware, high-speed data processing, and analytical algorithms that are combined to make data-based recommendations. As more and more data are collected, Watson has the capability to learn over time. In simple terms, according to IBM, Watson gathers hundreds of thousands of possible solutions from a huge data bank, evaluates them using analytical techniques, and proposes only the best solutions for consideration. Watson provides not just a single solution, but rather a range of good solutions with a confidence level for each.

For example, at a data center in Virginia, to the delight of doctors and patients, Watson is already being used to speed up the approval of medical procedures. Citibank is beginning to explore how to use Watson to better serve its customers, and cancer specialists at more than a dozen hospitals in North America are using Watson to assist with the diagnosis and treatment of patients.[1]

This book is concerned with data-driven decision making and the use of analytical approaches in the decision-making process. Three developments spurred recent explosive growth in the use of analytical methods in business applications. First, technological advances—such as improved point-of-sale scanner technology and the collection of data through e-commerce and social networks, data obtained by sensors on all kinds of mechanical devices such as aircraft engines, automobiles, and farm machinery through the so-called Internet of Things and data generated from personal electronic devices—produce incredible amounts of data for businesses. Naturally, businesses want to use these data to improve the efficiency and profitability of their operations, better understand their customers, price their products more effectively, and gain a competitive advantage. Second, ongoing research has resulted in numerous methodological developments, including advances in computational approaches to effectively handle and explore massive amounts of data, faster algorithms for optimization and simulation, and more effective approaches for visualizing data. Third, these methodological developments were paired with an explosion in computing power and storage capability. Better computing hardware, parallel computing, and, more recently, cloud computing (the remote use of hardware and software over the Internet) have enabled businesses to solve big problems more quickly and more accurately than ever before.

In summary, the availability of massive amounts of data, improvements in analytic methodologies, and substantial increases in computing power have all come together to result in a dramatic upsurge in the use of analytical methods in business and a reliance on the discipline that is the focus of this text: business analytics. As stated in the Preface, the purpose of this text is to provide students with a sound conceptual understanding of the role that business analytics plays in the decision-making process. To reinforce the applications orientation of the text and to provide a better understanding of the variety of applications in which analytical methods have been used successfully, Analytics in Action articles are presented throughout the book. Each Analytics in Action article summarizes an application of analytical methods in practice.

1.1 Decision Making

It is the responsibility of managers to plan, coordinate, organize, and lead their organizations to better performance. Ultimately, managers' responsibilities require that they make strategic, tactical, or operational decisions. **Strategic decisions** involve higher-level issues concerned with the overall direction of the organization; these decisions define the organization's overall goals and aspirations for the future. Strategic decisions are usually the domain of higher-level executives and have a time horizon of three to five years. **Tactical decisions** concern how the organization should achieve the goals and objectives set by its strategy, and they are usually the responsibility of midlevel management. Tactical decisions usually span a year and thus are revisited annually or even every six months. **Operational decisions** affect how the firm is run from day to day; they are the domain of operations managers, who are the closest to the customer.

Consider the case of the Thoroughbred Running Company (TRC). Historically, TRC had been a catalog-based retail seller of running shoes and apparel. TRC sales revenues grew quickly as it changed its emphasis from catalog-based sales to Internet-based sales. Recently, TRC decided that it should also establish retail stores in the malls and downtown areas of major cities. This strategic decision will take the firm in a new direction that it hopes will complement its Internet-based strategy. TRC middle managers will therefore have to make a variety of tactical decisions in support of this strategic decision, including

[1] "IBM's Watson Is Learning Its Way to Saving Lives," Fastcompany web site, December 8, 2012; H. Landi, "IBM Watson Health Touts Recent Studies Showing AI Improves How Physicians Treat Cancer," FierceHealthcare web site, June 4, 2019.

how many new stores to open this year, where to open these new stores, how many distribution centers will be needed to support the new stores, and where to locate these distribution centers. Operations managers in the stores will need to make day-to-day decisions regarding, for instance, how many pairs of each model and size of shoes to order from the distribution centers and how to schedule their sales personnel's work time.

Regardless of the level within the firm, *decision making* can be defined as the following process:

1. Identify and define the problem.
2. Determine the criteria that will be used to evaluate alternative solutions.
3. Determine the set of alternative solutions.
4. Evaluate the alternatives.
5. Choose an alternative.

Step 1 of decision making, identifying and defining the problem, is the most critical. Only if the problem is well-defined, with clear metrics of success or failure (step 2), can a proper approach for solving the problem (steps 3 and 4) be devised. Decision making concludes with the choice of one of the alternatives (step 5).

There are a number of approaches to making decisions: tradition ("We've always done it this way"), intuition ("gut feeling"), and rules of thumb ("As the restaurant owner, I schedule twice the number of waiters and cooks on holidays"). The power of each of these approaches should not be underestimated. Managerial experience and intuition are valuable inputs to making decisions, but what if relevant data were available to help us make more informed decisions? With the vast amounts of data now generated and stored electronically, it is estimated that the amount of data stored by businesses more than doubles every two years. How can managers convert these data into knowledge that they can use to be more efficient and effective in managing their businesses?

1.2 Business Analytics Defined

What makes decision making difficult and challenging? Uncertainty is probably the number one challenge. If we knew how much the demand will be for our product, we could do a much better job of planning and scheduling production. If we knew exactly how long each step in a project will take to be completed, we could better predict the project's cost and completion date. If we knew how stocks will perform, investing would be a lot easier.

Another factor that makes decision making difficult is that we often face such an enormous number of alternatives that we cannot evaluate them all. What is the best combination of stocks to help me meet my financial objectives? What is the best product line for a company that wants to maximize its market share? How should an airline price its tickets so as to maximize revenue?

*Some firms and industries use the simpler term, **analytics**. Analytics is often thought of as a broader category than business analytics, encompassing the use of analytical techniques in the sciences and engineering as well. In this text, we use **business analytics** and **analytics** synonymously.*

Business analytics is the scientific process of transforming data into insight for making better decisions.[2] Business analytics is used for data-driven or fact-based decision making, which is often seen as more objective than other alternatives for decision making.

As we shall see, the tools of business analytics can aid decision making by creating insights from data, by improving our ability to more accurately forecast for planning, by helping us quantify risk, and by yielding better alternatives through analysis and optimization. A study based on a large sample of firms that was conducted by researchers at MIT's Sloan School of Management and the University of Pennsylvania concluded that firms guided by data-driven decision making have higher productivity and market value and increased output and profitability.[3]

[2] We adopt the definition of analytics developed by the Institute for Operations Research and the Management Sciences (INFORMS).

[3] E. Brynjolfsson, L. M. Hitt, and H. H. Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" Thirty-Second International Conference on Information Systems, Shanghai, China, December 2011.

1.3 A Categorization of Analytical Methods and Models

Business analytics can involve anything from simple reports to the most advanced optimization techniques (methods for finding the best course of action). Analytics is generally thought to comprise three broad categories of techniques: descriptive analytics, predictive analytics, and prescriptive analytics.

Descriptive Analytics

Descriptive analytics encompasses the set of techniques that describes what has happened in the past. Examples are data queries, reports, descriptive statistics, data visualization including data dashboards, some data-mining techniques, and basic what-if spreadsheet models.

Appendix B, at the end of this book, describes how to use Microsoft Access to conduct data queries.

A **data query** is a request for information with certain characteristics from a database. For example, a query to a manufacturing plant's database might be for all records of shipments to a particular distribution center during the month of March. This query provides descriptive information about these shipments: the number of shipments, how much was included in each shipment, the date each shipment was sent, and so on. A report summarizing relevant historical information for management might be conveyed by the use of descriptive statistics (means, measures of variation, etc.) and data-visualization tools (tables, charts, and maps). Simple descriptive statistics and data-visualization techniques can be used to find patterns or relationships in a large database.

Data dashboards are collections of tables, charts, maps, and summary statistics that are updated as new data become available. Dashboards are used to help management monitor specific aspects of the company's performance related to their decision-making responsibilities. For corporate-level managers, daily data dashboards might summarize sales by region, current inventory levels, and other company-wide metrics; front-line managers may view dashboards that contain metrics related to staffing levels, local inventory levels, and short-term sales forecasts.

Data mining is the use of analytical techniques for better understanding patterns and relationships that exist in large data sets. For example, by analyzing text on social network platforms like Twitter, data-mining techniques (including cluster analysis and sentiment analysis) are used by companies to better understand their customers. By categorizing certain words as positive or negative and keeping track of how often those words appear in tweets, a company like Apple can better understand how its customers are feeling about a product like the Apple Watch.

Predictive Analytics

Predictive analytics consists of techniques that use models constructed from past data to predict the future or ascertain the impact of one variable on another. For example, past data on product sales may be used to construct a mathematical model to predict future sales. This mode can factor in the product's growth trajectory and seasonality based on past patterns. A packaged-food manufacturer may use point-of-sale scanner data from retail outlets to help in estimating the lift in unit sales due to coupons or sales events. Survey data and past purchase behavior may be used to help predict the market share of a new product. All of these are applications of predictive analytics.

Linear regression, time series analysis, some data-mining techniques, and simulation, often referred to as risk analysis, all fall under the banner of predictive analytics. We discuss all of these techniques in greater detail later in this text.

Data mining, previously discussed as a descriptive analytics tool, is also often used in predictive analytics. For example, a large grocery store chain might be interested in developing a targeted marketing campaign that offers a discount coupon on potato chips. By studying historical point-of-sale data, the store may be able to use data mining to predict which customers are the most likely to respond to an offer on discounted chips by purchasing higher-margin items such as beer or soft drinks in addition to the chips, thus increasing the store's overall revenue.

Simulation involves the use of probability and statistics to construct a computer model to study the impact of uncertainty on a decision. For example, banks often use simulation to model investment and default risk in order to stress-test financial models. Simulation is also often used in the pharmaceutical industry to assess the risk of introducing a new drug.

Prescriptive Analytics

Prescriptive analytics differs from descriptive and predictive analytics in that **prescriptive analytics** indicates a course of action to take; that is, the output of a prescriptive model is a decision. Predictive models provide a forecast or prediction, but do not provide a decision. However, a forecast or prediction, when combined with a rule, becomes a prescriptive model. For example, we may develop a model to predict the probability that a person will default on a loan. If we create a rule that says if the estimated probability of default is more than 0.6, we should not award a loan, now the predictive model, coupled with the rule is prescriptive analytics. These types of prescriptive models that rely on a rule or set of rules are often referred to as **rule-based models**.

Other examples of prescriptive analytics are portfolio models in finance, supply network design models in operations, and price-markdown models in retailing. Portfolio models use historical investment return data to determine which mix of investments will yield the highest expected return while controlling or limiting exposure to risk. Supply-network design models provide plant and distribution center locations that will minimize costs while still meeting customer service requirements. Given historical data, retail price markdown models yield revenue-maximizing discount levels and the timing of discount offers when goods have not sold as planned. All of these models are known as **optimization models**, that is, models that give the best decision subject to the constraints of the situation.

Another type of modeling in the prescriptive analytics category is **simulation optimization** which combines the use of probability and statistics to model uncertainty with optimization techniques to find good decisions in highly complex and highly uncertain settings. Finally, the techniques of **decision analysis** can be used to develop an optimal strategy when a decision maker is faced with several decision alternatives and an uncertain set of future events. Decision analysis also employs **utility theory**, which assigns values to outcomes based on the decision maker's attitude toward risk, loss, and other factors.

In this text we cover all three areas of business analytics: descriptive, predictive, and prescriptive. Table 1.1 shows how the chapters cover the three categories.

1.4 Big Data

On any given day, 500 million tweets and 294 billion e-mails are sent, 95 million photos and videos are shared on Instagram, 350 million photos are posted on Facebook, and 3.5 billion searches are made with Google.[4] It is through technology that we have truly been thrust into the data age. Because data can now be collected electronically, the available amounts of it are staggering. The Internet, cell phones, retail checkout scanners, surveillance video, and sensors on everything from aircraft to cars to bridges allow us to collect and store vast amounts of data in real time.

In the midst of all of this data collection, the term *big data* has been created. There is no universally accepted definition of big data. However, probably the most accepted and most general definition is that **big data** is any set of data that is too large or too complex to be handled by standard data-processing techniques and typical desktop software. IBM describes the phenomenon of big data through the four Vs: volume, velocity, variety, and veracity, as shown in Figure 1.1.[5]

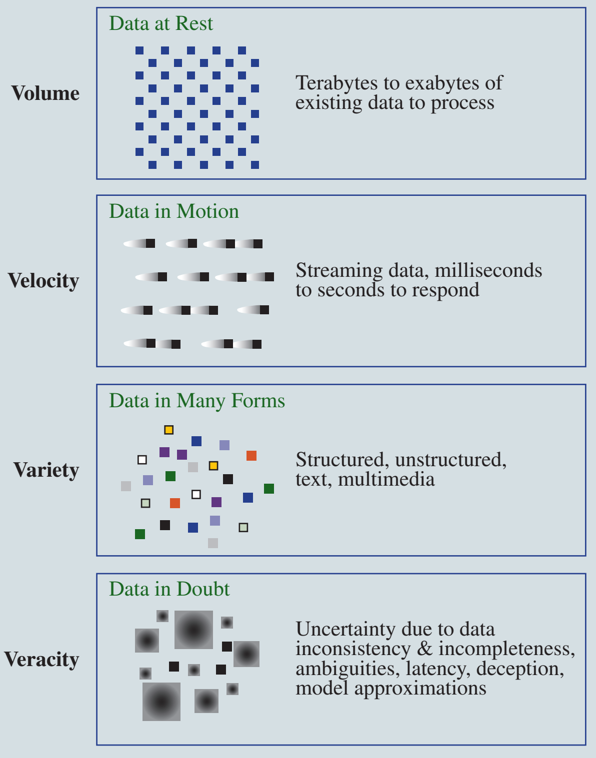
[4]J. Desjardins, "How Much Data Is Generated Each Day?" Visual Capitalist web site, April 15, 2019.

[5]IBM web site: www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg.

TABLE 1.1 Coverage of Business Analytics Topics in This Text

Chapter	Title	Descriptive	Predictive	Prescriptive
1	Introduction	●	●	●
2	Descriptive Statistics	●		
3	Data Visualization	●		
4	Probability: An Introduction to Modeling Uncertainty	●		
5	Descriptive Data Mining	●		
6	Statistical Inference	●		
7	Linear Regression		●	
8	Time Series and Forecasting		●	
9	Predictive Data Mining		●	
10	Spreadsheet Models	●	●	●
11	Monte Carlo Simulation		●	●
12	Linear Optimization Models			●
13	Integer Linear Optimization Models			●
14	Nonlinear Optimization Models			●
15	Decision Analysis			●

FIGURE 1.1 The Four Vs of Big Data



Source: IBM.

Volume

Because data are collected electronically, we are able to collect more of it. To be useful, these data must be stored, and this storage has led to vast quantities of data. Many companies now store in excess of 100 terabytes of data (a terabyte is 1,024 gigabytes).

Velocity

Real-time capture and analysis of data present unique challenges both in how data are stored, and the speed with which those data can be analyzed for decision making. For example, the New York Stock Exchange collects 1 terabyte of data in a single trading session, and having current data and real-time rules for trades and predictive modeling are important for managing stock portfolios.

Variety

In addition to the sheer volume and speed with which companies now collect data, more complicated types of data are now available and are proving to be of great value to businesses. Text data are collected by monitoring what is being said about a company's products or services on social media platforms such as Twitter. Audio data are collected from service calls (on a service call, you will often hear "this call may be monitored for quality control"). Video data collected by in-store video cameras are used to analyze shopping behavior. Analyzing information generated by these nontraditional sources is more complicated in part because of the processing required to transform the data into a numerical form that can be analyzed.

Veracity

Veracity has to do with how much uncertainty is in the data. For example, the data could have many missing values, which makes reliable analysis a challenge. Inconsistencies in units of measure and the lack of reliability of responses in terms of bias also increase the complexity of the data.

Businesses have realized that understanding big data can lead to a competitive advantage. Although big data represents opportunities, it also presents challenges in terms of data storage and processing, security, and available analytical talent.

The four Vs indicate that big data creates challenges in terms of how these complex data can be captured, stored, and processed; secured; and then analyzed. Traditional databases more or less assume that data fit into nice rows and columns, but that is not always the case with big data. Also, the sheer volume (the first V) often means that it is not possible to store all of the data on a single computer. This has led to new technologies like **Hadoop**—an open-source programming environment that supports big data processing through distributed storage and distributed processing on clusters of computers. Essentially, Hadoop provides a divide-and-conquer approach to handling massive amounts of data, dividing the storage and processing over multiple computers. **MapReduce** is a programming model used within Hadoop that performs the two major steps for which it is named: the map step and the reduce step. The map step divides the data into manageable subsets and distributes it to the computers in the cluster (often termed nodes) for storing and processing. The reduce step collects answers from the nodes and combines them into an answer to the original problem. Technologies like Hadoop and MapReduce, paired with relatively inexpensive computer power, enable cost-effective processing of big data; otherwise, in some cases, processing might not even be possible.

While some sources of big data are publicly available (Twitter, weather data, etc.), much of it is private information. Medical records, bank account information, and credit card transactions, for example, are all highly confidential and must be protected from computer hackers. **Data security**, the protection of stored data from destructive forces or unauthorized users, is of critical importance to companies. For example, credit card transactions are potentially very useful for understanding consumer behavior, but compromise of these data could lead to unauthorized use of the credit card or identity theft. A 2016 study of 383 companies in 12 countries conducted by the Ponemon Institute and IBM found that the average cost of a data breach is $3.86 million.[6] Companies such as Target, Anthem, JPMorgan Chase, Yahoo!, Facebook, Marriott, Equifax, and Home Depot have faced major data breaches costing millions of dollars.

The complexities of the 4 Vs have increased the demand for analysts, but a shortage of qualified analysts has made hiring more challenging. More companies are searching for **data scientists**, who know how to effectively process and analyze massive amounts of data because they are well trained in both computer science and statistics. Next we discuss three examples of how companies are collecting big data for competitive advantage.

Kroger Understands Its Customers[7] Kroger is the largest retail grocery chain in the United States. It sends over 11 million pieces of direct mail to its customers each quarter. The quarterly mailers each contain 12 coupons that are tailored to each household based on several years of shopping data obtained through its customer loyalty card program. By collecting and analyzing consumer behavior at the individual household level, and better matching its coupon offers to shopper interests, Kroger has been able to realize a far higher redemption rate on its coupons. In the six-week period following distribution of the mailers, over 70% of households redeem at least one coupon, leading to an estimated coupon revenue of $10 billion for Kroger.

MagicBand at Disney[8] The Walt Disney Company offers a wristband to visitors to its Orlando, Florida, Disney World theme park. Known as the MagicBand, the wristband contains technology that can transmit more than 40 feet and can be used to track each visitor's location in the park in real time. The band can link to information that allows Disney to better serve its visitors. For example, prior to the trip to Disney World, a visitor might be asked to fill out a survey on his or her birth date and favorite rides, characters, and restaurant table type and location. This information, linked to the MagicBand, can allow Disney employees using smartphones to greet you by name as you arrive, offer you products they know you prefer, wish you a happy birthday, have your favorite characters show up as you wait in line or have lunch at your favorite table. The MagicBand can be linked to your credit card, so there is no need to carry cash or a credit card. And during your visit, your movement throughout the park can be tracked and the data can be analyzed to better serve you during your next visit to the park.

General Electric and the Internet of Things[9] The **Internet of Things (IoT)** is the technology that allows data, collected from sensors in all types of machines, to be sent over the Internet to repositories where it can be stored and analyzed. This ability to collect data from products has enabled the companies that produce and sell those products to better serve their customers and offer new services based on analytics. For example, each day General Electric (GE) gathers nearly 50 million pieces of data from 10 million sensors on medical equipment and aircraft engines it has sold to customers throughout the world. In the case of aircraft engines, through a service agreement with its customers, GE collects data each time an airplane powered by its engines takes off and lands. By analyzing these data, GE can better predict when maintenance is needed, which helps customers avoid unplanned maintenance and downtime and helps ensure safe operation. GE can also use the data to better control how the plane is flown, leading to a decrease in fuel cost by flying more efficiently. GE spun off a new company called GE Digital 2.0 which operates as a stand-alone company focused on software that leverages IoT data. In 2018, GE announced that it would spin off a new company from its existing GE Digital business that will focus on industrial IoT applications.

Although big data is clearly one of the drivers for the strong demand for analytics, it is important to understand that, in some sense, big data issues are a subset of analytics. Many very valuable applications of analytics do not involve big data, but rather traditional data sets that are very manageable by traditional database and analytics software. The key to

---

[6] S. Shepard, "The Average Cost of a Data Breach," Security Today web site, July 17, 2018.

[7] Based on "Kroger Knows Your Shopping Patterns Better than You Do," Forbes.com, October 23, 2013.

[8] Based on "Disney's $1 Billion Bet on a Magical Wristband," Wired.com, March 10, 2015.

[9] Based on "G.E. Opens Its Big Data Platform," NYTimes.com, October 9, 2014; "GE Announces New Industrial IoT Software Business," Forbes web site, December 14, 2018.