

|                                                                       |          |
|-----------------------------------------------------------------------|----------|
| <i>Preface</i>                                                        | page xi  |
| <i>About the Author</i>                                               | xviii    |
| <b>Part I Conceptual Introductions</b>                                | <b>1</b> |
| <b>1 Introduction</b>                                                 | <b>3</b> |
| Online Datasets                                                       | 3        |
| 1.1 What Is Data Science?                                             | 3        |
| 1.2 Where Do We See Data Science?                                     | 6        |
| 1.2.1 Finance                                                         | 6        |
| 1.2.2 Public Policy                                                   | 7        |
| 1.2.3 Politics                                                        | 8        |
| 1.2.4 Healthcare                                                      | 9        |
| 1.2.5 Climate Change                                                  | 10       |
| 1.2.6 Urban Planning                                                  | 10       |
| 1.2.7 Education                                                       | 11       |
| 1.2.8 Libraries                                                       | 12       |
| 1.3 How Does Data Science Relate to Other Fields?                     | 13       |
| 1.3.1 Data Science and Statistics                                     | 13       |
| 1.3.2 Data Science and Computer Science                               | 14       |
| 1.3.3 Data Science and Engineering                                    | 14       |
| 1.3.4 Data Science and Business Analytics                             | 15       |
| 1.3.5 Data Science, Social Science, and Computational Social Science  | 15       |
| 1.4 The Relationship between Data Science and Information Science     | 16       |
| 1.4.1 Information vs. Data                                            | 17       |
| 1.4.2 Users in Information Science                                    | 17       |
| 1.4.3 Data Science in Information Schools (iSchools)                  | 18       |
| 1.5 The Relationship between Data Science and Artificial Intelligence | 18       |
| 1.6 Computational Thinking                                            | 20       |
| 1.7 Skills for Data Science                                           | 23       |
| 1.8 Tools for Data Science                                            | 29       |
| 1.9 Issues of Ethics, Bias, and Privacy in Data Science               | 30       |
| 1.9.1 Concerns for Data Users                                         | 31       |
| 1.9.2 Concerns for Data Scientists                                    | 32       |
| 1.9.3 Data Supply Chain                                               | 33       |
| 1.9.4 Bias and Inclusion                                              | 34       |
| 1.9.5 Considering Best Practices and Codes of Conduct                 | 34       |
| Summary                                                               | 35       |

|                                         |    |
|-----------------------------------------|----|
| Key Terms                               | 36 |
| Conceptual Questions                    | 37 |
| Hands-On Problems                       | 37 |
| <i>References</i>                       | 40 |
| <b>2 Data</b>                           | 43 |
| Online Datasets                         | 43 |
| 2.1 Introduction                        | 43 |
| 2.2 Data Types                          | 44 |
| 2.2.1 Structured Data                   | 44 |
| 2.2.2 Unstructured Data                 | 45 |
| 2.2.3 Challenges with Unstructured Data | 45 |
| 2.3 Data Collections                    | 46 |
| 2.3.1 Open Data                         | 46 |
| 2.3.2 Social Media Data                 | 47 |
| 2.3.3 Multimodal Data                   | 47 |
| 2.3.4 Synthetic Data                    | 48 |
| 2.4 Data Storage and Presentation       | 49 |
| 2.5 Data Pre-Processing                 | 54 |
| 2.5.1 Data Cleaning                     | 55 |
| 2.5.2 Data Integration                  | 57 |
| 2.5.3 Data Transformation               | 58 |
| 2.5.4 Data Reduction                    | 58 |
| 2.5.5 Data Discretization               | 59 |
| Summary                                 | 67 |
| Key Terms                               | 67 |
| Conceptual Questions                    | 68 |
| Hands-On Problems                       | 68 |
| Further Reading and Resources           | 73 |
| <i>References</i>                       | 73 |
| <b>3 Techniques</b>                     | 75 |
| Online Datasets                         | 75 |
| 3.1 Introduction                        | 75 |
| 3.2 Data Analysis and Data Analytics    | 76 |
| 3.3 Descriptive Analysis                | 77 |
| 3.3.1 Variables                         | 78 |
| 3.3.2 Frequency Distribution            | 81 |
| 3.3.3 Measures of Centrality            | 85 |
| 3.3.4 Dispersion of a Distribution      | 87 |
| 3.4 Diagnostic Analytics                | 91 |
| 3.4.1 Correlations                      | 92 |
| 3.5 Predictive Analytics                | 95 |
| 3.6 Prescriptive Analytics              | 96 |
| 3.7 Exploratory Analysis                | 97 |
| 3.8 Mechanistic Analysis                | 98 |
| 3.8.1 Regression                        | 98 |

|                                        |     |
|----------------------------------------|-----|
| Summary                                | 100 |
| Key Terms                              | 102 |
| Conceptual Questions                   | 103 |
| Hands-On Problems                      | 104 |
| Further Reading and Resources          | 107 |
| References                             | 107 |
| <b>Part II Tools for Data Science</b>  | 109 |
| <b>4 Introduction to R</b>             | 111 |
| 4.1 Introduction                       | 111 |
| 4.2 Getting Access to R                | 111 |
| 4.3 Getting Started with R             | 113 |
| 4.3.1 Basics                           | 113 |
| 4.3.2 Data Types                       | 115 |
| 4.3.3 Data Structures                  | 117 |
| 4.3.4 Control Structures               | 120 |
| 4.3.5 Functions                        | 123 |
| 4.4 R Markdown                         | 124 |
| 4.4.1 Using R Markdown                 | 124 |
| 4.4.2 Basic Syntax                     | 125 |
| Summary                                | 127 |
| Key Terms                              | 127 |
| Conceptual Questions                   | 127 |
| Hands-On Problems                      | 128 |
| Further Reading and Resources          | 129 |
| References                             | 130 |
| <b>5 R for Statistical Analysis</b>    | 131 |
| Online Datasets                        | 131 |
| 5.1 Introduction                       | 131 |
| 5.2 Importing Data                     | 132 |
| 5.3 Calculating Basic Statistics       | 133 |
| 5.4 Graphics and Data Visualization    | 135 |
| 5.4.1 Installing ggplot2               | 136 |
| 5.4.2 Loading the Data                 | 136 |
| 5.4.3 Plotting the Data                | 136 |
| 5.5 Statistical Inference              | 141 |
| 5.5.1 Correlation                      | 142 |
| 5.5.2 Hypothesis Testing               | 143 |
| 5.5.3 Comparing Means Using t-Test     | 144 |
| 5.5.4 Analysis of Variance using ANOVA | 147 |
| Summary                                | 148 |
| Key Terms                              | 149 |
| Conceptual Questions                   | 150 |
| Hands-On Problems                      | 150 |
| Further Reading and Resources          | 152 |

|                                                             |     |
|-------------------------------------------------------------|-----|
| <b>6 Cloud Computing</b>                                    | 153 |
| 6.1 Cloud Computing                                         | 153 |
| 6.2 Google Cloud Platform                                   | 154 |
| 6.2.1 Hadoop                                                | 157 |
| 6.3 Microsoft Azure                                         | 160 |
| 6.4 Amazon Web Services (AWS)                               | 165 |
| 6.5 Moving Between Cloud Platforms                          | 170 |
| 6.6 Summary                                                 | 171 |
| Key Terms                                                   | 172 |
| Conceptual Questions                                        | 172 |
| Hands-on Problems                                           | 172 |
| Further Reading and Resources                               | 173 |
| References                                                  | 173 |
| <b>Part III Machine Learning for Data Science</b>           | 175 |
| <b>7 Machine Learning Introduction and Regression</b>       | 177 |
| Online Datasets                                             | 177 |
| 7.1 Introduction                                            | 177 |
| 7.2 What Is Machine Learning?                               | 178 |
| 7.3 Regression                                              | 183 |
| 7.4 Gradient Descent                                        | 189 |
| 7.5 Considerations for Applying Machine Learning Techniques | 197 |
| Summary                                                     | 198 |
| Key Terms                                                   | 199 |
| Conceptual Questions                                        | 200 |
| Hands-On Problems                                           | 200 |
| Further Reading and Resources                               | 202 |
| References                                                  | 202 |
| <b>8 Supervised Learning</b>                                | 203 |
| Online Datasets                                             | 203 |
| 8.1 Introduction                                            | 204 |
| 8.2 Logistic Regression                                     | 205 |
| 8.3 Softmax Regression                                      | 213 |
| 8.4 Classification with kNN                                 | 216 |
| 8.5 Decision Tree                                           | 220 |
| 8.5.1 Decision Rule                                         | 225 |
| 8.5.2 Classification Rule                                   | 225 |
| 8.5.3 Association Rule                                      | 226 |
| 8.6 Random Forest                                           | 229 |
| 8.7 Naive Bayes                                             | 234 |
| 8.8 Support Vector Machine (SVM)                            | 240 |
| Summary                                                     | 247 |
| Key Terms                                                   | 248 |
| Conceptual Questions                                        | 249 |

|                                                            |     |
|------------------------------------------------------------|-----|
| Hands-On Problems                                          | 250 |
| Further Reading and Resources                              | 256 |
| <i>References</i>                                          | 256 |
| <b>9 Unsupervised Learning</b>                             | 257 |
| Online Datasets                                            | 257 |
| 9.1 Introduction                                           | 257 |
| 9.2 Agglomerative Clustering                               | 258 |
| 9.3 Divisive Clustering                                    | 262 |
| 9.4 Expectation Maximization (EM)                          | 266 |
| 9.5 Introduction to Reinforcement Learning                 | 274 |
| Summary                                                    | 277 |
| Key Terms                                                  | 279 |
| Conceptual Questions                                       | 279 |
| Hands-On Problems                                          | 279 |
| Further Reading and Resources                              | 282 |
| <i>References</i>                                          | 282 |
| <b>Part IV Applications, Evaluations, and Methods</b>      | 283 |
| <b>10 Data Collection, Experimentation, and Evaluation</b> | 285 |
| 10.1 Introduction                                          | 285 |
| 10.2 Data Collection Methods                               | 286 |
| 10.2.1 Surveys                                             | 286 |
| 10.2.2 Survey Question Types                               | 286 |
| 10.2.3 Survey Audience                                     | 288 |
| 10.2.4 Survey Services                                     | 289 |
| 10.2.5 Analyzing Survey Data                               | 290 |
| 10.2.6 Pros and Cons of Surveys                            | 290 |
| 10.2.7 Interviews and Focus Groups                         | 291 |
| 10.2.8 Why Do an Interview?                                | 291 |
| 10.2.9 Why Focus Groups?                                   | 292 |
| 10.2.10 Interview or Focus Group Procedure?                | 292 |
| 10.2.11 Analyzing Interview Data                           | 293 |
| 10.2.12 Pros and Cons of Interviews and Focus Groups       | 294 |
| 10.2.13 Log and Diary Data                                 | 294 |
| 10.2.14 User Studies in Lab and Field                      | 296 |
| 10.3 Picking Data Collection and Analysis Methods          | 297 |
| 10.3.1 Introduction to Quantitative Methods                | 298 |
| 10.3.2 Introduction to Qualitative Methods                 | 299 |
| 10.3.3 Mixed Method Studies                                | 300 |
| 10.4 Evaluation                                            | 301 |
| 10.4.1 Comparing Models                                    | 302 |
| 10.4.2 Training–Testing and A/B Testing                    | 304 |
| 10.4.3 Cross-Validation                                    | 306 |
| Summary                                                    | 307 |

|                                                            |     |
|------------------------------------------------------------|-----|
| Key Terms                                                  | 308 |
| Conceptual Questions                                       | 309 |
| Further Reading and Resources                              | 309 |
| <i>References</i>                                          | 309 |
| <b>11 Hands-On with Solving Data Problems</b>              | 311 |
| Online Datasets                                            | 311 |
| 11.1 Introduction                                          | 311 |
| 11.2 Collecting and Analyzing Reddit Data                  | 317 |
| 11.3 Collecting and Analyzing YouTube Data                 | 324 |
| 11.4 Analyzing Yelp Reviews and Ratings                    | 329 |
| Summary                                                    | 336 |
| Key Terms                                                  | 337 |
| Conceptual Questions                                       | 337 |
| Hands-On Problems                                          | 337 |
| <i>References</i>                                          | 340 |
| <i>Appendix A: Useful Formulas</i>                         | 341 |
| <i>Appendix B: Installing and Configuring Tools</i>        | 344 |
| B.1 R                                                      | 344 |
| B.2 RStudio                                                | 344 |
| <i>Appendix C: Using MySQL with R</i>                      | 347 |
| C.1 Getting Started with MySQL                             | 347 |
| C.1.1 Obtaining MySQL                                      | 347 |
| C.1.2 Logging in to MySQL                                  | 348 |
| C.2 Creating and Inserting Records                         | 350 |
| C.2.1 Importing Data                                       | 350 |
| C.2.2 Creating a Table                                     | 351 |
| C.2.3 Inserting Records                                    | 352 |
| C.3 Retrieving Records                                     | 352 |
| C.3.1 Reading Details about Tables                         | 353 |
| C.3.2 Retrieving Information from Tables                   | 353 |
| C.4 Searching in MySQL                                     | 355 |
| C.4.1 Searching within Field Values                        | 355 |
| C.4.2 Full-Text Searching with Indexing                    | 355 |
| C.5 Accessing MySQL with R                                 | 356 |
| <i>Appendix D: Introduction to Other Popular Databases</i> | 359 |
| D.1 NoSQL                                                  | 359 |
| D.2 MongoDB                                                | 359 |
| D.3 Google BigQuery                                        | 360 |
| <i>Appendix E: Data Science Jobs</i>                       | 361 |
| E.1 Marketing                                              | 362 |
| E.2 Corporate Retail and Sales                             | 362 |
| E.3 Legal                                                  | 363 |
| E.4 Health and Human Services                              | 364 |
| <i>Index</i>                                               | 366 |