A Hands-On Introduction to Data Science

This book introduces the field of data science in a practical and accessible manner, using a hands-on approach that assumes no prior knowledge of the subject. The foundational ideas and techniques of data science are provided independently from technology, allowing students to easily develop a firm understanding of the subject without a strong technical background, as well as being presented with material that will have continual relevance even after tools and technologies change. Using popular data science tools such as Python and R, the book offers many examples of real-life applications, with practice ranging from small to big data. A suite of online material for both instructors and students provides a strong supplement to the book, including datasets, chapter slides, solutions, sample exams, and curriculum suggestions. This entry-level textbook is ideally suited to readers from a range of disciplines wishing to build a practical, working knowledge of data science.

Chirag Shah is Associate Professor at University of Washington in Seattle. Before, he was a faculty member at Rutgers University, where he also served as the Coordinator of Data Science concentration for Master of Information. He has been teaching data science and machine learning courses to undergraduate, masters, and Ph.D. students for more than a decade. His research focuses on issues of search and recommendations using data mining and machine learning. Dr. Shah received his M.S. in Computer Science from the University of Massachusetts Amherst, and his Ph.D. in Information Science from the University of North Carolina Chapel Hill. He directs the InfoSeeking Lab, supported by awards from the National Science Foundation (NSF), the National Institute of Health (NIH), the Institute of Museum and Library Services (IMLS), as well as Amazon, Google, and Yahoo!. He was a Visiting Research Scientist at Spotify and has served as a consultant to the United Nations Data Analytics on various data science projects. He is currently working at Amazon in Seattle on large-scale e-commerce data and machine learning problems as Amazon Scholar.
A Hands-On Introduction to Data Science

CHIRAG SHAH
University of Washington
To my amazingly smart and sweet daughters – Sophie, Zoe, and Sarah – for adding colors and curiosity back to doing science and living life!
## Contents

*Preface*  
*About the Author*  
*Acknowledgments*  

**Part I: Conceptual Introductions**  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>xv</td>
</tr>
<tr>
<td>About the Author</td>
<td>xx</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>xxii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 What Is Data Science?</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Where Do We See Data Science?</td>
<td>5</td>
</tr>
<tr>
<td>1.2.1 Finance</td>
<td>6</td>
</tr>
<tr>
<td>1.2.2 Public Policy</td>
<td>7</td>
</tr>
<tr>
<td>1.2.3 Politics</td>
<td>8</td>
</tr>
<tr>
<td>1.2.4 Healthcare</td>
<td>9</td>
</tr>
<tr>
<td>1.2.5 Urban Planning</td>
<td>10</td>
</tr>
<tr>
<td>1.2.6 Education</td>
<td>10</td>
</tr>
<tr>
<td>1.2.7 Libraries</td>
<td>11</td>
</tr>
<tr>
<td>1.3 How Does Data Science Relate to Other Fields?</td>
<td>11</td>
</tr>
<tr>
<td>1.3.1 Data Science and Statistics</td>
<td>12</td>
</tr>
<tr>
<td>1.3.2 Data Science and Computer Science</td>
<td>13</td>
</tr>
<tr>
<td>1.3.3 Data Science and Engineering</td>
<td>13</td>
</tr>
<tr>
<td>1.3.4 Data Science and Business Analytics</td>
<td>14</td>
</tr>
<tr>
<td>1.3.5 Data Science, Social Science, and Computational Social Science</td>
<td>14</td>
</tr>
<tr>
<td>1.4 The Relationship between Data Science and Information Science</td>
<td>15</td>
</tr>
<tr>
<td>1.4.1 Information vs. Data</td>
<td>16</td>
</tr>
<tr>
<td>1.4.2 Users in Information Science</td>
<td>16</td>
</tr>
<tr>
<td>1.4.3 Data Science in Information Schools (iSchools)</td>
<td>17</td>
</tr>
<tr>
<td>1.5 Computational Thinking</td>
<td>17</td>
</tr>
<tr>
<td>1.6 Skills for Data Science</td>
<td>21</td>
</tr>
<tr>
<td>1.7 Tools for Data Science</td>
<td>27</td>
</tr>
<tr>
<td>1.8 Issues of Ethics, Bias, and Privacy in Data Science</td>
<td>29</td>
</tr>
<tr>
<td>Summary</td>
<td>30</td>
</tr>
<tr>
<td>Key Terms</td>
<td>31</td>
</tr>
<tr>
<td>Conceptual Questions</td>
<td>32</td>
</tr>
<tr>
<td>Hands-On Problems</td>
<td>32</td>
</tr>
</tbody>
</table>

© in this web service Cambridge University Press  
www.cambridge.org
2 Data
   2.1 Introduction 37
   2.2 Data Types 37
      2.2.1 Structured Data 38
      2.2.2 Unstructured Data 38
      2.2.3 Challenges with Unstructured Data 39
   2.3 Data Collections 39
      2.3.1 Open Data 40
      2.3.2 Social Media Data 41
      2.3.3 Multimodal Data 41
      2.3.4 Data Storage and Presentation 42
   2.4 Data Pre-processing 47
      2.4.1 Data Cleaning 48
      2.4.2 Data Integration 50
      2.4.3 Data Transformation 51
      2.4.4 Data Reduction 51
      2.4.5 Data Discretization 52
   Summary 59
   Key Terms 60
   Conceptual Questions 60
   Hands-On Problems 61
   Further Reading and Resources 65

3 Techniques
   3.1 Introduction 66
   3.2 Data Analysis and Data Analytics 67
   3.3 Descriptive Analysis 67
      3.3.1 Variables 68
      3.3.2 Frequency Distribution 71
      3.3.3 Measures of Centrality 75
      3.3.4 Dispersion of a Distribution 77
   3.4 Diagnostic Analytics 82
      3.4.1 Correlations 82
   3.5 Predictive Analytics 84
   3.6 Prescriptive Analytics 85
   3.7 Exploratory Analysis 86
   3.8 Mechanistic Analysis 87
      3.8.1 Regression 87
   Summary 89
   Key Terms 91
   Conceptual Questions 92
   Hands-On Problems 92
   Further Reading and Resources 95
Part II: Tools for Data Science

4 UNIX
4.1 Introduction
4.2 Getting Access to UNIX
4.3 Connecting to a UNIX Server
4.3.1 SSH
4.3.2 FTP/SCP/SFTP
4.4 Basic Commands
4.4.1 File and Directory Manipulation Commands
4.4.2 Process-Related Commands
4.4.3 Other Useful Commands
4.4.4 Shortcuts
4.5 Editing on UNIX
4.5.1 The vi Editor
4.5.2 The Emacs Editor
4.6 Redirections and Piping
4.7 Solving Small Problems with UNIX
Summary
Key Terms
Conceptual Questions
Hands-On Problems
Further Reading and Resources

5 Python
5.1 Introduction
5.2 Getting Access to Python
5.2.1 Download and Install Python
5.2.2 Running Python through Console
5.2.3 Using Python through Integrated Development Environment (IDE)
5.3 Basic Examples
5.4 Control Structures
5.5 Statistics Essentials
5.5.1 Importing Data
5.5.2 Plotting the Data
5.5.3 Correlation
5.5.4 Linear Regression
5.5.5 Multiple Linear Regression
5.6 Introduction to Machine Learning
5.6.1 What Is Machine Learning?
5.6.2 Classification (Supervised Learning)
5.6.3 Clustering (Unsupervised Learning)
5.6.4 Density Estimation (Unsupervised Learning)
<table>
<thead>
<tr>
<th>Chapter 6</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>6.2</td>
<td>Getting Access to R</td>
</tr>
<tr>
<td>6.3</td>
<td>Getting Started with R</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Basics</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Control Structures</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Functions</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Importing Data</td>
</tr>
<tr>
<td>6.4</td>
<td>Graphics and Data Visualization</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Installing ggplot2</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Loading the Data</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Plotting the Data</td>
</tr>
<tr>
<td>6.5</td>
<td>Statistics and Machine Learning</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Basic Statistics</td>
</tr>
<tr>
<td>6.5.2</td>
<td>Regression</td>
</tr>
<tr>
<td>6.5.3</td>
<td>Classification</td>
</tr>
<tr>
<td>6.5.4</td>
<td>Clustering</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 7</th>
<th>MySQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>7.2</td>
<td>Getting Started with MySQL</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Obtaining MySQL</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Logging into MySQL</td>
</tr>
<tr>
<td>7.3</td>
<td>Creating and Inserting Records</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Importing Data</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Creating a Table</td>
</tr>
<tr>
<td>7.3.3</td>
<td>Inserting Records</td>
</tr>
<tr>
<td>7.4</td>
<td>Retrieving Records</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Reading Details about Tables</td>
</tr>
<tr>
<td>7.4.2</td>
<td>Retrieving Information from Tables</td>
</tr>
<tr>
<td>7.5</td>
<td>Searching in MySQL</td>
</tr>
<tr>
<td>7.5.1</td>
<td>Searching within Field Values</td>
</tr>
<tr>
<td>7.5.2</td>
<td>Full-Text Searching with Indexing</td>
</tr>
</tbody>
</table>
### 7.6 Accessing MySQL with Python

Page 196

### 7.7 Accessing MySQL with R

Page 199

### 7.8 Introduction to Other Popular Databases

- **7.8.1 NoSQL**
- **7.8.2 MongoDB**
- **7.8.3 Google BigQuery**

Page 200

### Summary

Page 202

### Key Terms

Page 202

### Conceptual Questions

Page 203

### Hands-On Problems

Page 203

### Further Reading and Resources

Page 204

### Part III: Machine Learning for Data Science

Page 207

#### 8 Machine Learning Introduction and Regression

- **8.1 Introduction**
- **8.2 What Is Machine Learning?**
- **8.3 Regression**
- **8.4 Gradient Descent**

Page 209

### Summary

Page 229

### Key Terms

Page 230

### Conceptual Questions

Page 231

### Hands-On Problems

Page 231

### Further Reading and Resources

Page 233

#### 9 Supervised Learning

- **9.1 Introduction**
- **9.2 Logistic Regression**
- **9.3 Softmax Regression**
- **9.4 Classification with kNN**
- **9.5 Decision Tree**
  - **9.5.1 Decision Rule**
  - **9.5.2 Classification Rule**
  - **9.5.3 Association Rule**
- **9.6 Random Forest**
- **9.7 Naïve Bayes**
- **9.8 Support Vector Machine (SVM)**

Page 235

### Summary

Page 252

### Key Terms

Page 257

### Conceptual Questions

Page 257

### Hands-On Problems

Page 257

### Further Reading and Resources

Page 260

Page 266

Page 272

Page 279

Page 280

Page 281

Page 288
10 Unsupervised Learning 290
10.1 Introduction 290
10.2 Agglomerative Clustering 291
10.3 Divisive Clustering 295
10.4 Expectation Maximization (EM) 299
10.5 Introduction to Reinforcement Learning 309
Summary 312
Key Terms 313
Conceptual Questions 314
Hands-On Problems 314
Further Reading and Resources 317

Part IV: Applications, Evaluations, and Methods 319

11 Hands-On with Solving Data Problems 321
11.1 Introduction 321
11.2 Collecting and Analyzing Twitter Data 328
11.3 Collecting and Analyzing YouTube Data 336
11.4 Analyzing Yelp Reviews and Ratings 342
Summary 349
Key Terms 350
Conceptual Questions 350
Practice Questions 351

12 Data Collection, Experimentation, and Evaluation 354
12.1 Introduction 354
12.2 Data Collection Methods 355
  12.2.1 Surveys 355
  12.2.2 Survey Question Types 355
  12.2.3 Survey Audience 357
  12.2.4 Survey Services 358
  12.2.5 Analyzing Survey Data 359
  12.2.6 Pros and Cons of Surveys 360
  12.2.7 Interviews and Focus Groups 360
  12.2.8 Why Do an Interview? 360
  12.2.9 Why Focus Groups? 361
  12.2.10 Interview or Focus Group Procedure 361
  12.2.11 Analyzing Interview Data 362
  12.2.12 Pros and Cons of Interviews and Focus Groups 362
  12.2.13 Log and Diary Data 363
  12.2.14 User Studies in Lab and Field 364
12.3 Picking Data Collection and Analysis Methods 366
  12.3.1 Introduction to Quantitative Methods 366
12.3.2 Introduction to Qualitative Methods 368
12.3.3 Mixed Method Studies 369
12.4 Evaluation 370
  12.4.1 Comparing Models 370
  12.4.2 Training-Testing and A/B Testing 372
  12.4.3 Cross-Validation 374
Summary 376
Key Terms 377
Conceptual Questions 377
Further Reading and Resources 378

Appendices
Appendix A: Useful Formulas from Differential Calculus 379
  Further Reading and Resources 380
Appendix B: Useful Formulas from Probability 381
  Further Reading and Resources 381
Appendix C: Useful Resources 383
  C.1 Tutorials 383
  C.2 Tools 383
Appendix D: Installing and Configuring Tools 385
  D.1 Anaconda 385
  D.2 IPython (Jupyter) Notebook 385
  D.3 Spyder 387
  D.4 R 387
  D.5 RStudio 388
Appendix E: Datasets and Data Challenges 390
  E.1 Kaggle 390
  E.2 RecSys 391
  E.3 WSDM 391
  E.4 KDD Cup 392
Appendix F: Using Cloud Services 393
  F.1 Google Cloud Platform 394
  F.2 Hadoop 398
  F.3 Microsoft Azure 400
  F.4 Amazon Web Services (AWS) 403
Appendix G: Data Science Jobs 407
  G.1 Marketing 408
  G.2 Corporate Retail and Sales 409
  G.3 Legal 409
  G.4 Health and Human Services 410
## Contents

**Appendix H: Data Science and Ethics**

- H.1 Data Supply Chain ................................................................. 412
- H.2 Bias and Inclusion ............................................................... 414
- H.3 Considering Best Practices and Codes of Conduct ................. 414

**Appendix I: Data Science for Social Good** ................................... 416

**Index** ......................................................................................... 418
Preface

Data science is one of the fastest-growing disciplines at the university level. We see more job postings that require training in data science, more academic appointments in the field, and more courses offered, both online and in traditional settings. It could be argued that data science is nothing novel, but just statistics through a different lens. What matters is that we are living in an era in which the kind of problems that could be solved using data are driving a huge wave of innovations in various industries – from healthcare to education, and from finance to policy-making. More importantly, data and data analysis are playing an increasingly large role in our day-to-day life, including in our democracy. Thus, knowing the basics of data and data analysis has become a fundamental skill that everyone needs, even if they do not want to pursue a degree in computer science, statistics, or data science. Recognizing this, many educational institutions have started developing and offering not just degrees and majors in the field but also minors and certificates in data science that are geared toward students who may not become data scientists but could still benefit from data literacy skills in the same way every student learns basic reading, writing, and comprehension skills.

This book is not just for data science majors but also for those who want to develop their data literacy. It is organized in a way that provides a very easy entry for almost anyone to become introduced to data science, but it also has enough fuel to take one from that beginning stage to a place where they feel comfortable obtaining and processing data for deriving important insights. In addition to providing basics of data and data processing, the book teaches standard tools and techniques. It also examines implications of the use of data in areas such as privacy, ethics, and fairness. Finally, as the name suggests, this text is meant to provide a hands-on introduction to these topics. Almost everything presented in the book is accompanied by examples and exercises that one could try – sometimes by hand and other times using the tools taught here. In teaching these topics myself, I have found this to be a very effective method.

The remainder of this preface explains how this book is organized, how it could be used for fulfilling various teaching needs, and what specific requirements a student needs to meet to make the most out of it.

Requirements and Expectations

This book is intended for advanced undergraduates or graduate students in information science, computer science, business, education, psychology, sociology, and related fields...
who are interested in data science. It is not meant to provide in-depth treatment of any
programming language, tool, or platform. Similarly, while the book covers topics such as
machine learning and data mining, it is not structured to give detailed theoretical instruction
on them; rather, these topics are covered in the context of applying them to solving various
data problems with hands-on exercises.

The book assumes very little to no prior exposure to programming or technology. It does,
however, expect the student to be comfortable with computational thinking (see Chapter 1)
and the basics of statistics (covered in Chapter 3). The student should also have general
computer literacy, including skills to download, install, and configure software, do file
operations, and use online resources. Each chapter lists specific requirements and expecta-
tions, many of which can be met by going over some other parts of the book (usually an
earlier chapter or an appendix).

Almost all the tools and software used in this book are free. There is no requirement of
a specific operating system or computer architecture, but it is assumed that the student has
a relatively modern computer with reasonable storage, memory, and processing power. In
addition, a reliable and preferably high-speed Internet connection is required for several
parts of this book.

Structure of the Book

The book is organized in four parts. Part I includes three chapters that serve as the
foundations of data science. Chapter 1 introduces the field of data science, along with
various applications. It also points out important differences and similarities with related
fields of computer science, statistics, and information science. Chapter 2 describes the
nature and structure of data as we encounter today. It initiates the student about data
formats, storage, and retrieval infrastructures. Chapter 3 introduces several important
techniques for data science. These techniques stem primarily from statistics and include
correlation analysis, regression, and introduction to data analytics.

Part II of this book includes chapters to introduce various tools and platforms such as
UNIX (Chapter 4), Python (Chapter 5), R (Chapter 6), and MySQL (Chapter 7). It is
important to keep in mind that, since this is not a programming or database book, the
objective here is not to go systematically into various parts of these tools. Rather, we focus
on learning the basics and the relevant aspects of these tools to be able to solve various data
problems. These chapters therefore are organized around addressing various data-driven
problems. In the chapters covering Python and R, we also introduce basic machine learning.

But machine learning is a crucial topic for data science that cannot be treated just as an
afterthought, which is why Part III of this book is devoted to it. Specifically, Chapter 8
provides a more formal introduction to machine learning and includes a few techniques that
are basic and broadly applicable at the same time. Chapter 9 describes in some depth
supervised learning methods, and Chapter 10 presents unsupervised learning. It should be
noted that, since this book is focused on data science and not core computer science or
mathematics, we skip much of the underlying math and formal structuring while discussing
and applying machine learning techniques. The chapters in Part III, however, do present machine learning methods and techniques using adequate math in order to discuss the theories and intuitions behind them in detail.

Finally, Part IV of this book takes the techniques from Part I, as well as the tools from Parts II and III to start applying them to problems of real-life significance. In Chapter 11, we take this opportunity by applying various data science techniques to several real-life problems, including those involving social media, finance, and social good. Finally, Chapter 12 provides additional coverage into data collection, experimentation, and evaluation.

The book is full of extra material that either adds more value and knowledge to your existing data science theories and practices, or provides broader and deeper treatment of some of the topics. Throughout the book, there are several FYI boxes that provide important and relevant information without interrupting the flow of the text, allowing the student to be aware of various issues such as privacy, ethics, and fairness without being overwhelmed by them. The appendices of this book provide quick reference to various formulations relating to differential calculus and probability, as well as helpful pointers and instructions for installing and configuring various tools used in the book. For those interested in using cloud-based platforms and tools, there is also an appendix that shows how to sign up, configure, and use them. Another appendix provides listing of various sources for obtaining small to large datasets for more practice and even participate in data challenges to win some cool prizes and recognition. There is also an appendix that provides helpful information related to data science jobs in various fields and what skills one should have to target those calls. Finally, a couple of appendices introduce the ideas of data ethics and data science for social good to inspire you to be a responsible and socially aware data citizen.

The book also has an online appendix (OA), accessible through the book’s website at www.cambridge.org/shah, which is regularly updated to reflect any changes in data and other resources. The primary purpose for this online appendix is to provide you with the most current and updated datasets or links to datasets that you can download and use in the dozens of examples and try-it-yourself exercises in the chapters, as well as data problems at the end of the chapters. Look for the icon at various places that inform you that you need to find the needed resource from OA. In the description of that exercise, you will see the specific number (e.g., OA 3.2) that tells you where exactly you should go in the online appendix.

**Using This Book in Teaching**

The book is quite deliberately organized around teaching data science to beginner computer science (CS) students or intermediate to advanced non-CS students. The book is modular, making it easier for both students and teachers to cover topics to the desired depth. This makes it quite suitable for the book to be used as a main reference book or textbook for
a data science curriculum. The following is a suggested curriculum path in data science using this book. It contains five courses, each lasting a semester or a quarter.

- Introduction to data science: Chapters 1 and 2, with some elements from Part II as needed.
- Data analytics: Chapter 3, with some elements from Part II as needed.
- Problem solving with data or programming for data science: Chapters 4–7.
- Research methods for data science: Chapter 12, with appropriate elements from Chapter 3 and Part II.

At the website for this book is a Resources tab with a section labeled “For Instructors.” This section contains sample syllabi for various courses that could be taught using this book, PowerPoint slides for each chapter, and other useful resources such as sample midterm and final exams. These resources make it easier for someone teaching this course for the first time to adapt the text as needed for his or her own data science curriculum.

Each chapter also has several conceptual questions and hands-on problems. The conceptual questions could be used in either in-class discussions, for homework, or for quizzes. For each new technique or problem covered in this book, there are at least two hands-on problems. One of these could be used in the class and the other one could be given for homework or exam. Most hands-on exercises in chapters are also immediately followed by hands-on homework exercises that a student could try for further practice, or an instructor could assign as homework or in-class practice assignment.

**Strengths and Unique Features of This Book**

Data science has a very visible presence these days, and it is not surprising that there are currently several available books and much material related to the field. *A Hands-On Introduction to Data Science* is different from the other books in several ways.

- It is targeted to students with very basic experience with technology. Students who fit in that category are majoring in information science, business, psychology, sociology, education, health, cognitive science, and indeed any area in which data can be applied. The study of data science should not be limited to those studying computer science or statistics. This book is intended for those audiences.
- The book starts by introducing the field of data science without any prior expectation of knowledge on the part of the reader. It then introduces the reader to some foundational ideas and techniques that are independent of technology. This does two things: (1) it provides an easier access point for a student without strong technical background; and (2) it presents material that will continue to be relevant even when tools and technologies change.
- Based on my own teaching and curriculum development experiences, I have found that most data science books on the market are divided into two categories: they are either too technical, making them suitable only for a limited audience, or they are structured to be...
simply informative, making it hard for the reader to actually use and apply data science tools and techniques. *A Hands-On Introduction to Data Science* is aimed at a nice middle ground: On the one hand, it is not simply describing data science, but also teaching real hands-on tools (Python, R) and techniques (from basic regression to various forms of machine learning). On the other hand, it does not require students to have a strong technical background to be able to learn and practice data science.

- *A Hands-On Introduction to Data Science* also examines implications of the use of data in areas such as privacy, ethics, and fairness. For instance, it discusses how unbalanced data used without enough care with a machine learning technique could lead to biased (and often unfair) predictions. There is also an introduction to the newly formulated General Data Protection Regulations (GDPR) in Europe.

- The book provides many examples of real-life applications, as well as practices ranging from small to big data. For instance, Chapter 4 has an example of working with housing data where simple UNIX commands could extract valuable insights. In Chapter 5, we see how multiple linear regression can be easily implemented using Python to learn how advertising spending on various media (TV, radio) could influence sales. Chapter 6 includes an example that uses R to analyze wines to predict which ones are of high quality. Chapters 8–10 on machine learning have many real-life and general interest problems from different fields as the reader is introduced to various techniques. Chapter 11 has hands-on exercises for collecting and analyzing social media data from services such as Twitter and YouTube, as well as working with large datasets (Yelp data with more than a million records). Many of the examples can be worked by hand or with everyday software, without requiring specialized tools. This makes it easier for a student to grasp a concept without having to worry about programming structures. This allows the book to be used for non-majors as well as professional certificate courses.

- Each chapter has plenty of in-chapter exercises where I walk the reader through solving a data problem using a new technique, homework exercises to do more practice, and more hands-on problems (often using real-life data) at the end of the chapters. There are 37 hands-on solved exercises, 46 hands-on try-it-yourself exercises, and 55 end-of-chapter hands-on problems.

- The book is supplemented by a generous set of material for instructors. These instructor resources include curriculum suggestions (even full-length syllabuses for some courses), slides for each chapter, datasets, program scripts, answers and solutions to each exercise, as well as sample mid-term exams and final projects.
Dr. Chirag Shah is Associate Professor at University of Washington in Seattle. Before, he was a faculty member at Rutgers University. He is a Senior Member of the Association for Computing Machinery (ACM). He received his Ph.D. in Information Science from University of North Carolina at Chapel Hill and a M.S. in Computer Science from the University of Massachusetts at Amherst.

His research interests include studies of interactive information seeking and retrieval, with applications to personalization and recommendation, as well as applying machine learning and data mining techniques to both big data and tiny data problems. He has published several books and peer-reviewed articles in the areas of information seeking and social media. He has developed Coagmento system for collaborative and social searching, IRIS (Information Retrieval and Interaction System) for investigating and implementing interactive IR activities, as well as several systems for collecting and analyzing data from social media channels, including award winning ContextMiner, InfoExtractor, TubeKit, and SOCRATES. He directs the InfoSeeking Lab, where he investigates issues related to information seeking, social media, and neural information retrieval. These research projects are supported by grants from the National Science Foundation (NSF), the National Institute of Health (NIH), the Institute of Museum and Library Services (IMLS), Amazon, Google, and Yahoo!. He also serves as a consultant to the United Nations Data Analytics on various data science projects involving social and political issues, peacekeeping, climate change, and energy. He spent his last sabbatical at Spotify as a Visiting Research Scientist and is currently consulting to Amazon on personalization and recommendation problems as an Amazon Scholar.

Dr. Shah has taught extensively to both undergraduate and graduate (masters and Ph.D.) students on topics of data science, machine learning, information retrieval (IR), human–computer interaction (HCI), and quantitative research methods. He has
also delivered special courses and tutorials at various international venues, and created massive open online courses (MOOCs) for platforms such as Coursera. He has developed several courses and curricula for data science and advised dozens of undergraduate and graduate students pursuing data science careers. This book is a result of his many years of teaching, advising, researching, and realizing the need for such a resource.

chirags@uw.edu
http://chiragshah.org
@chirag_shah
A book like this does not happen without a lot of people’s help and it would be rude of me to not acknowledge at least some of those people here.

As is the case with almost all of my projects, this one would not have been possible without the love and support of my wife Lori. She not only understands late nights and long weekends working on a project like this, but also keeps me grounded on what matters the most in life – my family, my students, and the small difference that I am trying to make in this world through the knowledge and skills I have.

My sweet and smart daughters – Sophie, Zoe, and Sarah – have also kept me connected to the reality while I worked on this book. They have inspired me to look beyond data and information to appreciate the human values behind them. After all, why bother doing anything in this book if it is not helping human knowledge and advancement in some way? I am constantly amazed by my kids’ curiosity and sense of adventure, because those are the qualities one needs in doing any kind of science, and certainly data science. A lot of the analyses and problem solving presented in this book fall under this category, where we are not simply processing some data, but are driven by a sense of curiosity and a quest to derive new knowledge.

This book, as I have noted in the Preface, happened organically over many years through developing and teaching various data science classes. And so I need to thank all of those students who sat in my classes or joined online, went through my material, asked questions, provided feedback, and helped me learn more. With every iteration of every class I have taught in data science, things have gotten better. In essence, what you are holding in your hands is the result of the best iteration so far.

In addition to hundreds (or thousands, in the case of MOOCs) of students over the years, there are specific students and assistants I need to thank for their direct and substantial contributions to this book. My InfoSeeking Lab assistants Liz Smith and Catherine McGowan have been tremendously helpful in not only proofreading, but also helping with literature review and contributing several pieces of writings. Similarly, Dongho Choi and Soumik Mandal, two of my Ph.D. students, have contributed substantially to some of the writings and many of the examples and exercises presented in this book. If it was not for the help and dedication of these four people, this book would have been delayed by at least a year.

I am also thankful to my Ph.D. students Souvick Ghosh, who provided some writeup on misinformation, and Ruoyuan Gao, for contributing to the topic of fairness and bias.

Finally, I am eternally grateful to the wonderful staff of Cambridge University Press for guiding me through the development of this book from the beginning. I would specifically call out Lauren Cowles, Lisa Pinto, and Stefanie Seaton. They have been an amazing team
helping me in almost every aspect of this book, ensuring that it meets the highest standards of quality and accessibility that one would expect from the Press. Writing a book is often a painful endeavor, but when you have a support team like this, it becomes possible and even a fun project!

I am almost certain that I have forgotten many more people to thank here, but they should know that it was a result of my forgetfulness and not ungratefulness.