

A Hands-On Introduction to Data Science with Python

Students will develop a practical understanding of data science with this hands-on textbook for introductory courses. This new edition is fully revised and updated, with numerous exercises and examples in the popular data science tool Python, a new chapter on using Python for statistical analysis, and a new chapter that demonstrates how to use Python within a range of cloud platforms. The many practice examples, drawn from real-life applications, range from small to big data and come to life in a new end-to-end project in Chapter 11. New “Data Science in Practice” boxes highlight how the concepts that have been introduced work within an industry context and many chapters include new sections on AI and Generative AI. A suite of online material for instructors provides a strong supplement to the book, including lecture slides, solutions, additional assessment material, and curriculum suggestions. Datasets and code are available for students online. This entry-level textbook is ideal for readers from a range of disciplines wishing to build a practical, working knowledge of data science.

Chirag Shah is Professor of Information and Computer Science at University of Washington (UW) in Seattle. He is the Founding Director for InfoSeeking Lab and Founding Co-Director of the Center for Responsibility in AI Systems & Experiences (RAISE). His research focuses on building, auditing, and correcting intelligent information access systems. Dr. Shah is a Distinguished Member of ACM as well as ASIS&T, and a Senior Member of IEEE. He has published nearly 200 peer-reviewed articles and authored several books, including textbooks on data science and machine learning. He regularly engages with industrial research labs at Amazon, ByteDance, Microsoft Research, and Spotify.

A Hands-On Introduction to Data Science

with Python

SECOND EDITION

CHIRAG SHAH

University of Washington



CAMBRIDGE
UNIVERSITY PRESS



CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/highereducation/isbn/9781009588928

DOI: 10.1017/9781009588911

© Chirag Shah 2026

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press & Assessment.

When citing this work, please include a reference to the DOI 10.1017/9781009588911

First published 2020

Second edition 2026

Cover image: (c) Paul Starosta / Stone / Getty Images

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data

Names: Shah, Chirag, author.

Title: A hands-on introduction to data science. With Python / Chirag Shah.

Description: Second edition. | New York : Cambridge University Press, 2025. | Includes bibliographical references and index.

Identifiers: LCCN 2025040975 (print) | LCCN 2025040976 (ebook) | ISBN 9781009588928 (hardback) | ISBN 9781009588942 (paperback) | ISBN 9781009588911 (epub)

Subjects: LCSH: Computer science—Textbooks. | Information technology—Textbooks. | Python (Computer program language)—Textbooks.

Classification: LCC QA76 .S693 2025 (print) | LCC QA76 (ebook)

LC record available at <https://lcn.loc.gov/2025040975>

LC ebook record available at <https://lcn.loc.gov/2025040976>

ISBN 978-1-009-58892-8 Hardback

ISBN 978-1-009-58894-2 Paperback

Additional resources for this publication at www.cambridge.org/shah-python2e

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

For EU product safety concerns, contact us at Calle de José Abascal, 56, 1º, 28003 Madrid, Spain, or email eugpsr@cambridge.org

Contents

<i>Preface</i>	<i>page xi</i>
<i>About the Author</i>	<i>xvii</i>
Part I Conceptual Introductions	1
1 Introduction	3
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
References	40
2 Data	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
References	73
3 Techniques	75
Online Appendix	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
Part II Tools for Data Science	109
4 Python	111
4.1 Introduction	111
4.2 Getting Access to Python	111
4.2.1 Download and Install Python	112
4.2.2 Running Python through Console	112
4.2.3 Using Python through Integrated Development Environment (IDE)	112
4.3 Basic Examples	115
4.4 Data Structures	117
4.5 Control Structures	120
4.6 Functions	122
4.7 Making Python Interactive	124
4.8 Installing and Using Python Packages	125
Summary	127
Key Terms	128
Conceptual Questions	128
Hands-On Problems	129
Further Reading and Resources	130
References	131
5 Python for Statistical Analysis	132
Online Datasets	132
5.1 Introduction	132
5.2 Statistics Essentials	133
5.3 Graphics and Data Visualization	136
5.3.1 Importing Data	136
5.3.2 Plotting the Data	136
5.4 Statistical Inference	137
5.4.1 Correlation	138
5.4.2 Hypothesis Testing	139
5.4.3 Comparing Means Using t-Test	140
5.4.4 Analysis of Variance Using ANOVA	143
Summary	145
Key Terms	146
Conceptual Questions	146
Hands-On Problems	147
Further Reading and Resources	149

6 Cloud Computing	151
6.1 Cloud Computing	151
6.2 Google Cloud Platform	152
6.2.1 Hadoop	155
6.3 Microsoft Azure	161
6.4 Amazon Web Services (AWS)	169
6.5 Moving between Cloud Platforms	176
Summary	177
Key Terms	177
Conceptual Questions	178
Hands-on Problems	178
References	179
 Part III Machine Learning for Data Science	 181
7 Machine Learning Introduction and Regression	183
Online Datasets	183
7.1 Introduction	183
7.2 What Is Machine Learning?	184
7.3 Regression	189
7.4 Gradient Descent	195
7.5 Considerations for Applying Machine Learning Techniques	203
Summary	205
Key Terms	206
Conceptual Questions	206
Hands-On Problems	207
Further Reading and Resources	208
References	209
 8 Supervised Learning	 210
Online Datasets	210
8.1 Introduction	211
8.2 Logistic Regression	212
8.3 Softmax Regression	221
8.4 Classification with kNN	225
8.5 Decision Tree	230
8.5.1 Decision Rule	235
8.5.2 Classification Rule	236
8.5.3 Association Rule	236
8.6 Random Forest	239
8.7 Naive Bayes	245
8.8 Support Vector Machine (SVM)	252
Summary	260
Key Terms	261
Conceptual Questions	262
Hands-On Problems	262

Further Reading and Resources	268
References	269
9 Unsupervised Learning	270
Online Datasets	270
9.1 Introduction	270
9.2 Agglomerative Clustering	271
9.3 Divisive Clustering	275
9.4 Expectation Maximization (EM)	279
9.5 Introduction to Reinforcement Learning	289
Summary	294
Key Terms	295
Conceptual Questions	296
Hands-On Problems	296
Further Reading and Resources	298
References	299
Part IV Applications, Evaluations, and Methods	301
10 Data Collection, Experimentation, and Evaluation	303
10.1 Introduction	303
10.2 Data Collection Methods	304
10.2.1 Surveys	304
10.2.2 Survey Question Types	304
10.2.3 Survey Audience	306
10.2.4 Survey Services	307
10.2.5 Analyzing Survey Data	308
10.2.6 Pros and Cons of Surveys	308
10.2.7 Interviews and Focus Groups	309
10.2.8 Why Do an Interview?	309
10.2.9 Why Focus Groups?	310
10.2.10 Interview or Focus Group Procedure?	310
10.2.11 Analyzing Interview Data	311
10.2.12 Pros and Cons of Interviews and Focus Groups	312
10.2.13 Log and Diary Data	312
10.2.14 User Studies in Lab and Field	314
10.3 Picking Data Collection and Analysis Methods	315
10.3.1 Introduction to Quantitative Methods	316
10.3.2 Introduction to Qualitative Methods	317
10.3.3 Mixed Method Studies	318
10.4 Evaluation	319
10.4.1 Comparing Models	320
10.4.2 Training–Testing and A/B Testing	322
10.4.3 Cross-Validation	324
Summary	325
Key Terms	326

Conceptual Questions	327
Further Reading and Resources	327
References	327
11 Hands-On with Solving Data Problems	329
Online Datasets	329
11.1 Introduction	329
11.2 Collecting and Analyzing Reddit Data	336
11.3 Collecting and Analyzing YouTube Data	342
11.4 Analyzing Yelp Reviews and Ratings	349
Summary	355
Key Terms	356
Conceptual Questions	356
Hands-On Problems	356
References	359
<i>Appendix A: Useful Formulas</i>	360
<i>Appendix B: Installing and Configuring Tools</i>	363
B.1 Anaconda	363
B.2 IPython (Jupyter) Notebook	363
B.3 Spyder	363
<i>Appendix C: Using MySQL with Python</i>	366
C.1 Getting Started with MySQL	366
C.1.1 Obtaining MySQL	366
C.1.2 Logging in to MySQL	367
C.2 Creating and Inserting Records	369
C.2.1 Importing Data	369
C.2.2 Creating a Table	370
C.2.3 Inserting Records	371
C.3 Retrieving Records	371
C.3.1 Reading Details about Tables	372
C.3.2 Retrieving Information from Tables	372
C.4 Searching in MySQL	373
C.4.1 Searching within Field Values	374
C.4.2 Full-Text Searching with Indexing	374
C.5 Accessing MySQL with Python	375
<i>Appendix D: Introduction to Other Popular Databases</i>	378
D.1 NoSQL	378
D.2 MongoDB	378
D.3 Google BigQuery	379
<i>Appendix E: Data Science Jobs</i>	380
E.1 Marketing	381
E.2 Corporate Retail and Sales	381
E.3 Legal	382
E.4 Health and Human Services	383
<i>Index</i>	385

Preface

Data science, along with artificial intelligence (AI) and machine learning (ML), continues to be one of the fastest-growing disciplines at education institutions and in industry. We see more job postings that require training in data science, AI, and ML, more academic appointments in these fields, 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. However, what matters is that we are living in an era in which the kind of problems that could be solved using data, AI, and ML are driving a huge wave of innovations in various industries – from healthcare to education, and from finance to policymaking. More importantly, data, AI, and ML are playing an increasingly large role in our day-to-day life, including in our democracy. Thus, knowing the basics of these areas 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 institutes have been developing and offering not just degrees and majors in these fields but also minors and certificates in data science, AI, and ML that are geared toward students who may not become data scientists but could still benefit from data literacy skills in the same way that 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 a reader 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 over many years, 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 in order 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 an 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 expectations, 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 for 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 from and similarities to the related fields of computer science, statistics, and information science. Chapter 2 describes the nature and structure of the data encountered 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 an introduction to data analytics.

Part II of this book includes chapters to introduce various necessary tools and platforms. Chapters 4 and 5 cover Python programming language – from basics to its applications in statistical analysis. What you see in these chapters can be done on your laptop but at times, and certainly in the industry, you are likely to need cloud-based infrastructure, for various reasons including security, compliance, and scalability. Therefore, in Chapter 6, we work through three of the most popular cloud platforms and see how they can be leveraged for doing data science with Python. 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 the tools to be able to solve various data problems. These chapters therefore are organized around addressing various data-driven problems.

Solving problems with data-driven approaches can hardly be done without machine learning. This is such a crucial topic for data science that it cannot be

treated just as an afterthought, which is why Part III of this book is devoted to it. Specifically, Chapter 7 provides a more formal introduction to machine learning and includes a few techniques that are basic and broadly applicable at the same time. Chapter 8 describes in some depth supervised learning methods, and Chapter 9 presents unsupervised learning. It should be noted that since this book is focused on data science and not on 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, and starts applying them to problems of real-life significance. We begin in Chapter 10 by reviewing methods for data collection, experimentation, and evaluation. This sets us up for going beyond just what is covered in this book and thinking broadly about how one might address real-life data science problems. Finally, we go all-in with our hands-on approach in Chapter 11, where we work on four different case studies of end-to-end data science projects involving healthcare, social media, finance, and social good.

The book is full of extra material that either adds more value and knowledge to your existing data science theories and practices or provides a 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. Two appendices of the book provide a 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. Another appendix teaches how to use MySQL, a popular database system, to use with R for data accessing and processing. 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.

As a new feature of this second edition, there are several boxes labeled “DS in Practice” that present ideas, guidance, and real-world examples of how the concepts and exercises covered in that chapter are applied in industry. Also, to make things more streamlined, all the datasets from hands-on examples, try-it-yourself problems, and hands-on problems are available to download from the book’s website at www.cambridge.org/shah-python2e. Look for a list of these datasets at the beginning of each chapter that informs you the specific online dataset (OD) that you will need to download. In the description of that exercise, you will see a specific number (e.g., OD 3.2) that tells you where exactly you should go in the OD.

The book also has an online appendix (OA), accessible through the same website, which is regularly updated to reflect any changes in data, resources, and links. The primary purpose of this online appendix is to provide you with the most current and updated pointers and information. The combination of the printed and the online material through the OD and OA serves as a unique resource for aspiring data scientists and instructors that has both comprehensiveness and recency.

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 the book quite suitable for use 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 4–6, with relevant project(s) from Chapter 11.
- Machine learning for data science: Chapters 7–9.
- Research methods for data science: Chapter 10, with appropriate elements from Chapter 3 and Part II.

On 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 mid-term 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 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 could be given for homework or an 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 as an 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 with Python* is different from other books in several ways.

- It is targeted at students with very basic experience of technology. Students who fit into that category are majoring in information science, business, psychology, sociology, education, health, cognitive science, or 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 a strong technical background; and (2) it presents material that will continue being relevant even when tools and technologies change.
- On the basis of 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 with Python* is aimed at a nice middle ground. On the one hand, the book does not simply describe data science but also teaches real hands-on tools (Python, cloud computing) 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 with Python* 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.
- The book provides many examples of real-life applications, as well as practices ranging from small to big data. For instance, Chapter 3 has an example of understanding oxygen uptake and expired ventilation. In Chapter 7, we see how multiple linear regression can be easily implemented using Python to learn how ambience, food, and service features could influence restaurant ratings. Chapter 8 has a hands-on example that shows how to predict, using customer data from a bank, whether the client will subscribe a term deposit. Chapters 7–9 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 Reddit 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 39 hands-on solved exercises, 40 try-it-yourself exercises, and 57 end of chapter hands-on problems. To help with theory homework or exams, there are also 62 conceptual questions at the end of the chapters.
- The book is supplemented by a generous set of materials for instructors. These instructor resources include curriculum suggestions (even full-length syllabi 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.

What’s New in the Second Edition

- NEW: Added information boxes “DS in Practice” in almost every chapter to highlight how the data science theory and practice presented in the book gets used in industry.
- NEW: Additional hands-on examples and problems, as well as conceptual questions.
- NEW: All datasets are available to download right from the book’s website. These datasets are now clearly listed at the start of each chapter under a newly added “Online Datasets” section.
- NEW: Chapter on “Python for Statistical Analysis” that goes further into using Python for statistical inference (e.g., hypothesis testing).
- NEW: Chapter on cloud computing that shows how to use Python for doing data science through popular cloud platforms – AWS, Azure, and Google Cloud. These are essential for working in industry and large-scale projects.
- NEW: An end-to-end project with Reddit data in Chapter 10.
- NEW: Several sections at appropriate places to reflect the influence of generative AI and AI in general. Example: the section on synthetic data in Chapter 2.
- NEW: A detailed discussion on data ethics for data users and data scientists in Chapter 1.
- NEW: Several appendices that are now available online through the book’s website to allow for regular updates and provision of the most current and accurate information to supplement the book’s material.
- UPDATES: Several screenshots, numbers, and links. Some of the descriptions and instructions for problems have also been updated.
- UPDATES: Incorporated fixes for the errors found before and reported in the errata document.
- UPDATES: The chapters on UNIX and MySQL have been moved to an appendix.

About the Author



Dr. Chirag Shah is a Professor in the Information School (iSchool) at the University of Washington (UW) in Seattle. He is also an Adjunct Professor with the Paul G. Allen School of Computer Science & Engineering as well as the Human Centered Design & Engineering (HCDE) department. Before UW, he was at Rutgers University. He is the Founding Director for InfoSeeking Lab and Founding Co-Director of Center for Responsibility in AI Systems & Experiences (RAISE). Dr. Shah is a Research Associate at the University of Pretoria in South Africa and a Visiting Professor at Peking University in China. He received his Ph.D. in

Information Science from the University of North Carolina (UNC) at Chapel Hill and an MS in Computer Science from the University of Massachusetts (UMass) at Amherst.

Dr. Shah is a Distinguished Member of the Association for Computing Machinery (ACM) as well as the Association for Information Science & Technology (ASIS&T), as well as a Senior Member of Institute of Electrical and Electronics Engineers (IEEE).

Dr. Shah's research focuses on building, auditing, and correcting intelligent information access systems. In addition to creating AI-driven information access systems that provide more personalized reactive and proactive recommendations, he is also focusing on making such systems transparent, fair, and free of bias. He applies techniques from data science and machine learning for most of this work. He has published nearly 200 peer-reviewed articles and several books, including a textbook titled *A Hands-On Approach to Machine Learning* by Cambridge University Press. His research is supported by grants from National Science Foundation (NSF), Institute of Museum and Library Services (IMLS), Amazon, Google, and Yahoo!. He has served as a consultant to the United Nations Data Analytics on various data science projects involving social and political issues, peacekeeping, climate change, and energy.

Dr. Shah has taught extensively both undergraduate and graduate (Masters and Ph.D.) students on the 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 MOOCs (massive open online courses) 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.

He spent his sabbatical in 2018 at Spotify working on voice-based search and recommendation problems. In 2019, as an Amazon Scholar, he worked with Amazon’s Personalization team on applications involving personalized and task-oriented recommendations. In 2021, he visited Getty Images to work on improving their search platform with embedding-based deep semantic approaches. In 2020, he was a Visiting Researcher at Microsoft Research (MSR) AI and worked on an intelligent task manager. He returned to MSR in 2023 and 2024 to work on generative AI problems, specifically focusing on building, deploying, and evaluating AI agents.

chirag.shah@acm.org
<https://chiragshah.org>