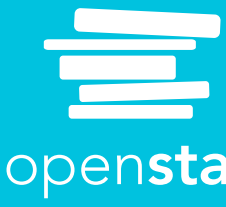


Principles of

Data Science

Principles of Data Science

SENIOR CONTRIBUTING AUTHORS

Dr. Shaun V. Ault, Valdosta State University

Dr. Soohyun Nam Liao, University of California San Diego

Larry Musolino, Pennsylvania State University

openstax™

OpenStax
Rice University
6100 Main Street MS-375
Houston, Texas 77005

To learn more about OpenStax, visit https://openstax.org.
Individual print copies and bulk orders can be purchased through our website.

COLOR PAPERBACK BOOK ISBN-13	**979-8-3851-6185-0**
B&W PAPERBACK BOOK ISBN-13	**979-8-3851-6186-7**
DIGITAL VERSION ISBN-13	**978-1-961584-60-0**
ORIGINAL PUBLICATION YEAR	**2025**

1 2 3 4 5 6 7 8 9 10 CJP 25

OpenStax

OpenStax provides free, peer-reviewed, openly licensed textbooks for introductory college and Advanced Placement® courses and low-cost, personalized courseware that helps students learn. A nonprofit ed tech initiative based at Rice University, we're committed to helping students access the tools they need to complete their courses and meet their educational goals.

Rice University

OpenStax is an initiative of Rice University. As a leading research university with a distinctive commitment to undergraduate education, Rice University aspires to path-breaking research, unsurpassed teaching, and contributions to the betterment of our world. It seeks to fulfill this mission by cultivating a diverse community of learning and discovery that produces leaders across the spectrum of human endeavor.

Philanthropic Support

OpenStax is grateful for the generous philanthropic partners who advance our mission to improve educational access and learning for everyone. To see the impact of our supporter community and our most updated list of partners, please visit openstax.org/foundation.

Arnold Ventures
Chan Zuckerberg Initiative
Chegg, Inc.
Arthur and Carlyse Ciocca Charitable Foundation
Digital Promise
Ann and John Doerr
Bill & Melinda Gates Foundation
Girard Foundation
Google Inc.
The William and Flora Hewlett Foundation
The Hewlett-Packard Company
Intel Inc.
Rusty and John Jaggers
The Calvin K. Kazanjian Economics Foundation
Charles Koch Foundation
Leon Lowenstein Foundation, Inc.
The Maxfield Foundation
Burt and Deedee McMurtry
Michelson 20MM Foundation
National Science Foundation
The Open Society Foundations
Jumee Yhu and David E. Park III
Brian D. Patterson USA-International Foundation
The Bill and Stephanie Sick Fund
Steven L. Smith & Diana T. Go
Stand Together
Robin and Sandy Stuart Foundation
The Stuart Family Foundation
Tammy and Guillermo Treviño
Valhalla Charitable Foundation
White Star Education Foundation
Schmidt Futures
William Marsh Rice University

CONTENTS

UNIT 4 MAINTAINING A PROFESSIONAL AND ETHICAL DATA SCIENCE PRACTICE

Preface

About OpenStax

OpenStax is part of Rice University, which is a 501(c)(3) nonprofit charitable corporation. As an educational initiative, it's our mission to improve educational access and learning for everyone. Through our partnerships with philanthropic organizations and our alliance with other educational resource companies, we're breaking down the most common barriers to learning. Because we believe that everyone should and can have access to knowledge.

About OpenStax Resources

Customization

Principles of Data Science is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY NC-SA) license, which means that you can non-commercially distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors, under the same license.

Because our books are openly licensed, you are free to use the entire book or select only the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. Visit the Instructor Resources section of your book page on OpenStax.org for more information.

Art Attribution

In *Principles of Data Science*, most art contains attribution to its title, creator or rights holder, host platform, and license within the caption. Because the art is openly licensed, non-commercial users or organizations may reuse the art as long as they provide the same attribution to its original source. (Commercial entities should contact OpenStax to discuss reuse rights and permissions.) To maximize readability and content flow, some art does not include attribution in the text. If you reuse art from this text that does not have attribution provided, use the following attribution: Copyright Rice University, OpenStax, under CC BY-NC-SA 4.0 license.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. In addition, the wide range of topics, data, technologies, and legal circumstances in data science change frequently, and portions of the textbook may become out of date. Since our books are web-based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on OpenStax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past and pending errata changes on your book page on OpenStax.org.

Format

You can access this textbook for free in web view or PDF through OpenStax.org, and for a low cost in print. The web view is the recommended format because it is the most accessible – including being WCAG 2.2 AA compliant – and most current. Print versions are available for individual purchase, or they may be ordered through your campus bookstore.

About *Principles of Data Science*

Summary

Principles of Data Science is intended as introductory material for a one- or two-semester course on data science. It is appropriate for undergraduate students interested in the rapidly growing field of data science;

this may include data science majors, data science minors, or students concentrating in business, finance, health care, engineering, the sciences, or a number of other fields where data science has become critically important. The material is designed to prepare students for future coursework and career applications in a data science–related field. It does not assume significant prior coding experience, nor does it assume completion of more than college algebra. The text provides foundational statistics instruction for students who may have a limited statistical background.

Coverage and Scope

Principles of Data Science emphasizes the use of Python code in relevant data science applications. Python provides a versatile programming language with libraries and frameworks for data manipulation, analysis, and machine learning. The book begins with an introduction to Python and presents Python libraries, algorithms, and functions as they are needed throughout. In occasional, focused instances, the authors also use Excel to illustrate the basic manipulation of data using functions, formulas, and tools for calculations, visualization, and financial analysis. R, a programming language used most often for statistical modeling, is briefly described and then summarized and applied to relevant examples in a book appendix. Excel and Python summaries are also provided in appendices at the end of the book.

The table of contents (TOC) is divided into ten chapters, organized in four units, intuitively following the standard data science cycle. The four units are:

Unit 1: *Introducing Data Science and Data Collection*
Unit 2: *Analyzing Data Using Statistics*
Unit 3: *Predicting and Modeling Using Data*
Unit 4: *Maintaining a Professional and Ethical Data Science Practice*

The learning objectives and curriculum of introductory data science courses vary, so this textbook aims to provide broader and more detailed coverage than an average single-semester course. Instructors can choose which chapters or sections they want to include in their particular course.

To enable this flexibility, chapters in this text can be used in a self-contained manner, although most chapters do cross-reference sections and chapters that precede or follow. More importantly, the authors have taken care to build topics gradually, from chapter to chapter, so instructors should bear this in mind when considering alternate sequence coverage.

Unit 1: Introducing Data Science and Data Collection starts off with **Chapter 1**'s explanation of the data science cycle (data collection and preparation, data analysis, and data reporting) and its practical applications in fields such as medicine, engineering, and business. **Chapter 1** also describes various types of datasets and provides the student with basic data summary tools from the Python `pandas` library. **Chapter 2** describes the processes of data collection and cleaning and the challenges of managing large datasets. It previews some of the qualitative ethical considerations that Chapters 7 and 8 later expand on.

Unit 2: Analyzing Data Using Statistics forms a self-contained unit that instructors may assign on a modular, more optional basis, depending on students' prior coursework. **Chapter 3** focuses on measures of center, variation, and position, leading up to probability theory and illustrating how to use Python with binomial and normal distributions. **Chapter 4** goes deeper into statistical analysis, demonstrating how to use Python to calculate confidence intervals, conduct hypothesis tests, and perform correlation and regression analysis.

The three chapters in ***Unit 3: Predicting and Modeling Using Data*** form the core of the book. **Chapter 5** introduces students to the concept and practical applications of time series. Chapter 5 provides focused examples of both Python and Excel techniques useful in forecasting time series, analyzing seasonality, and identifying measures of error. **Chapter 6** starts with distinguishing supervised vs. unsupervised machine learning and then develops some common methods of data classification, including logistic regression, clustering algorithms, and decision trees. Chapter 6 includes Python techniques for more sophisticated

statistical analyses such as regression with bootstrapping and multivariable regression. Finally, Chapter 6 refers back to the topics of data mining and big data introduced in Chapter 2.

Chapter 7 is a pedagogically rich chapter, with a balance of quantitative and qualitative content, covering the role of neural networks in deep learning and applications in large language models. The first four sections discuss the topics of neural networks (standard, recurrent, and convolutional), backpropagation, and deep learning. The real-life application of classifying handwritten numerals is used as an example. The last section dives into the important and rapidly changing technology of natural language processing (NLP), large language models (LLMs), and artificial intelligence (AI). While in-depth coverage of these evolving subjects is beyond the scope of this textbook, the pros/cons, the examples from technical and artistic applications, and the online resources provided in this section all serve as a good starting point for classroom discussion. This topic also naturally segues into the broader professional responsibility discussed in **Chapter 8**.

The final chapters in ***Unit 4: Maintaining a Professional and Ethical Data Science Practice*** help the student apply and adjust the specific techniques learned in the previous chapters to the real-life data analysis, decision-making, and communication situations they will encounter professionally. **Chapter 8** emphasizes the importance of ethical considerations along each part of the cycle: data collection; data preparation, analysis, and modeling; and reporting and visualization. Coverage of the issues in this chapter makes students aware of the subjective and sensitive aspects of privacy and informed consent at every step in the process. At the professional level, students learn more about the evolving standards for the relatively new field of data science, which may differ among industries or between the United States and other countries.

Chapter 9 circles back to some of the statistical concepts introduced in Chapters 3 and 4, with an emphasis on clear visual analysis of data trends. Chapter 9 provides a range of Python techniques to create boxplots and histograms for univariate data; to create line charts and trend curves for time series; to graph binomial, Poisson, and normal distributions; to generate heatmaps from geospatial data; and to create correlation heatmaps from multidimensional data.

Chapter 10 brings the student back to the practical decision-making setting introduced in Chapter 1. Chapter 10 helps the student address how to tailor the data analysis and presentation to the audience and purpose, how to validate the assumptions of a model, and how to write an effective executive summary.

The four **Appendices (A–D)** provide a practical set of references for Excel commands and commands for R statistical software as well as Python commands and algorithms. **Appendix A** uses a baseball dataset from Chapter 1 to illustrate basic Microsoft® Excel® software commands for manipulating, analyzing, summarizing, and graphing data. **Appendix B** provides a brief overview of data analysis with the open-source statistical computing package R (https://openstax.org/r/project), using a stock price example. **Appendix C** lists the approximately 60 Python algorithms used in the textbook, and **Appendix D** lists the code and syntax for the approximately 75 Python functions demonstrated in the textbook. Both **Appendices C and D** are organized in a tabular format, in consecutive chapter order, hyperlinked to the first significant use of each Python algorithm and function. (Instructors may find Appendices C and D especially useful in developing their teaching plan.)

Pedagogical Foundation

Because this is a practical, introductory-level textbook, math equations and code are presented clearly throughout. Particularly in the core chapters, students are introduced to key mathematical concepts, equations, and formulas, often followed by numbered Example Problems that encourage students to apply the concepts just covered in a variety of situations. Technical illustrations and Python code feature boxes build on and supplement the theory. Students are encouraged to try out the Python code from the feature boxes in the Google Colaboratory (https://openstax.org/r/colabresearch) (Colab) platform.

The authors have included a diverse mix of data types and sources for analysis, illustration, and discussion purposes. Some scenarios are fictional and/or internal to standard Python libraries, while other datasets come from external, real-world sources, both corporate and government (such as Federal Reserve Economic Data

(FRED) (https://openstax.org/r/nasdaq1), Statista (https://openstax.org/r/statista1), and Nasdaq (https://openstax.org/r/nasdaq)). Most scenarios are either summarized in an in-line table, or have datasets provided in a downloadable student spreadsheet for import as a .CSV file (Chapter 1 also discusses the .JSON format) and/or with a hyperlink to the external source. Some examples focus on scientific topics (e.g., the "classic" Iris flower dataset, annual temperature changes), while other datasets reflect phenomena with more nuanced socioeconomic issues (gender-based salary differences, cardiac disease markers in patients).

While the book's foundational chapters illustrate practical "techniques and tools," the more process-oriented chapters iteratively build on and emphasize an underlying framework of professional, responsible, and ethical data science practice. Chapter 1 refers the student to several national and international data science organizations that are developing professional standards. Chapter 2 emphasizes avoiding bias in survey and sample design. Chapter 8 discusses relevant privacy legislation. For further class exploration, Chapters 7 and 8 include online resources on mitigating bias and discrimination in machine learning models, including related Python libraries such as HolisticAI (https://openstax.org/r/holistic) and Fairlens (https://openstax.org/r/projectfairlens). Chapter 10 references several executive dashboards that support transparency in government.

The Group Projects at the end of each chapter encourage students to apply the techniques and considerations covered in the book using either datasets already provided or new data sources that they might receive from their instructors or in their own research. For example, project topics include the following: collecting data on animal extinction due to global warming (Chapter 2), predicting future trends in stock market prices (Chapter 5), diagnosing patients for liver disease (Chapter 7), and analyzing the severity of ransomware attacks (Chapter 8).

Key Features

The key in-chapter features, depending on chapter content and topics, may include the following:

- Learning Outcomes (LOs) to guide the student's progress through the chapter
- Example Problems, demonstrating calculations and solutions in-line
- Python code boxes, providing sample input code for and output from Google Colab
- Note boxes providing instructional tips to help with the practical aspects of the math and coding
- Data tables from a variety of social science and industry settings
- Technical charts and heatmaps to visually demonstrate code output and variable relationships
- Exploring Further boxes, with additional resources and online examples to extend learning
- Mathematical formulas and equations
- Links to downloadable spreadsheet containing key datasets referenced in the chapter for easy manipulation of data

End-of-chapter (EOC) elements, depending on chapter content and topics, may include the following:

- Key Terms
- Group Projects
- Chapter Review Questions
- Critical Thinking Questions
- Quantitative Problems

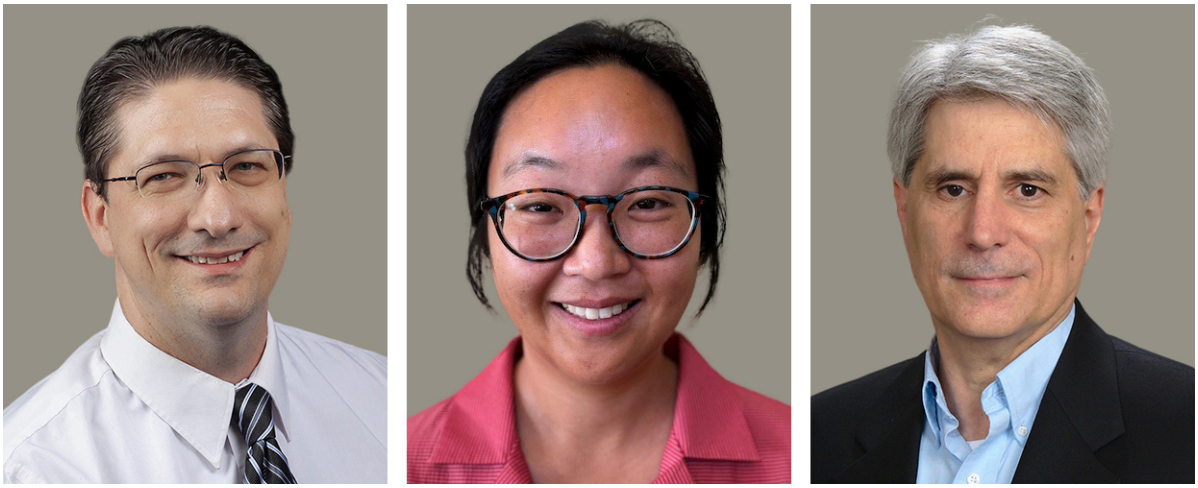
Answers [and Solutions] to Questions in the Book

The student-facing Answer Key at the end of the book provides the correct answer letter and text for Chapter Review questions (multiple-choice). An Instructor Solution Manual (ISM) will be available for verified instructors and downloadable from the restricted OpenStax Instructor Resources web page, with detailed solutions to Quantitative Problems, sample answers for Critical Thinking questions, and a brief explanation of the correct answer for Chapter Review questions. (Sample calculations, tables, code, or figures may be included, as

applicable.) An excerpt of the ISM, consisting of the solutions/sample answers for the odd-numbered questions only, will also be available as a Student Solution Manual (SSM), downloadable from the public OpenStax Student Resources web page. (Answers to the Group Projects are not provided, as they are integrative, exploratory, open-ended assignments.)

About the Authors

Senior Contributing Authors



Senior Contributing Authors (left to right): Shaun V. Ault, Soohyun Nam Liao, Larry Musolino

Dr. Shaun V. Ault, Valdosta State University. Dr. Ault joined the Valdosta State University faculty in 2012, serving as Department Head of Mathematics from 2017 to 2023 and Professor since 2021. He holds a PhD in mathematics from The Ohio State University, a BA in mathematics from Oberlin College, and a Bachelor of Music from the Oberlin Conservatory of Music. He previously taught at Fordham University and The Ohio State University. He is a Certified Institutional Review Board Professional and holds membership in the Mathematical Association of America, American Mathematical Society, and Society for Industrial and Applied Mathematics. He has research interests in algebraic topology and computational mathematics and has published in a number of peer-reviewed journal publications. He has authored two textbooks: *Understanding Topology: A Practical Introduction* (Johns Hopkins University Press, 2018) and, with Charles Kicey, *Counting Lattice Paths Using Fourier Methods. Applied and Numerical Harmonic Analysis*, Springer International (2019).

Dr. Soohyun Nam Liao, University of California San Diego. Dr. Liao joined the UC San Diego faculty in 2015, serving as Assistant Teaching Professor since 2021. She holds PhD and MS degrees in computer science and engineering from UC San Diego and a BS in electronics engineering from Seoul University, South Korea. She previously taught at Princeton University and was an engineer at Qualcomm Inc. She focuses on computer science (CS) education research as a means to support diversity and equity (DEI) in CS programs. Among her recent co-authored papers is, with Yunyi She, Korena S. Klimczak, and Michael E. Levin, “ClearMind Workshop: An ACT-Based Intervention Tailored for Academic Procrastination among Computing Students,” *SIGCSE* (1) 2024: 1216-1222. She has received a National Science Foundation grant to develop a toolkit for A14All (data science camps for high school students).

Larry Musolino, Pennsylvania State University. Larry Musolino joined the Penn State, Lehigh Valley, faculty in 2015, serving as Assistant Teaching Professor of Mathematics since 2022. He received an MS in mathematics from Texas A&M University, a MS in statistics from Rochester Institute of Technology (RIT), and MS degrees in computer science and in electrical engineering, both from Lehigh University. He received his BS in electrical engineering from City College of New York (CCNY). He previously was a Distinguished Member of Technical Staff in semiconductor manufacturing at LSI Corporation. He is a member of the Penn State OER (Open Educational Resources) Advisory Group and has authored a calculus open-source textbook. In addition, he co-authored an open-source *Calculus for Engineering* workbook. He has contributed to several OpenStax

textbooks, authoring the statistics chapters in the *Principles of Finance* textbook and editing and revising *Introductory Statistics, 2e,* and *Introductory Business Statistics, 2e*.

The authors wish to express their deep gratitude to Developmental Editor Ann West for her skillful editing and gracious shepherding of this manuscript. The authors also thank Technical Editor Dhawani Shah (PhD Statistics, Gujarat University) for contributing technical reviews of the chapters throughout the content development process.

Contributing Authors
Wisam Bukaita, Lawrence Technological University
Aeron Zentner, Coastline Community College

Reviewers
Wisam Bukaita, Lawrence Technological University
Drew Lazar, Ball State University
J. Hathaway, Brigham Young University-Idaho
Salvatore Morgera, University of South Florida
David H. Olsen, Utah Tech University
Thomas Pfaff, Ithaca College
Jian Yang, University of North Texas
Aeron Zentner, Coastline Community College

Additional Resources

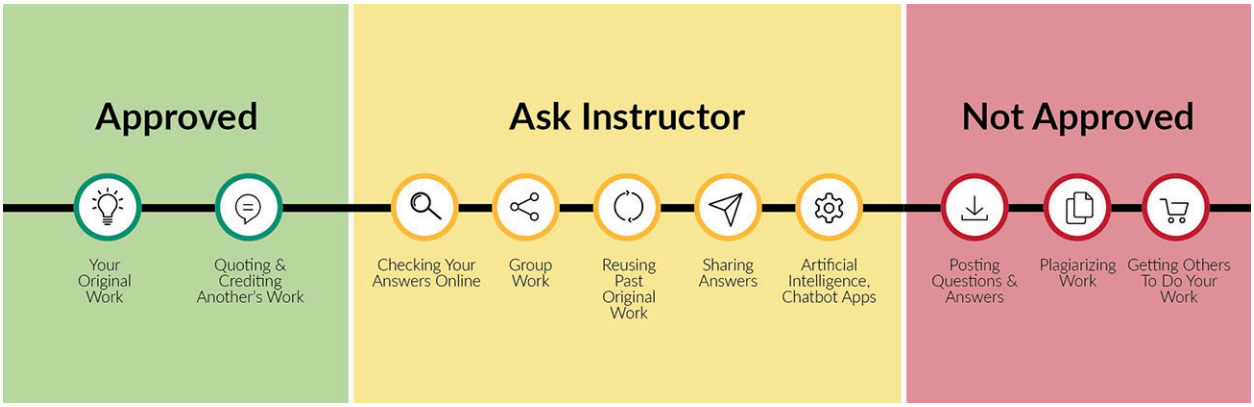
Student and Instructor Resources

We've compiled additional resources for both students and instructors, including Getting Started Guides. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

Academic Integrity

Academic integrity builds trust, understanding, equity, and genuine learning. While students may encounter significant challenges in their courses and their lives, doing their own work and maintaining a high degree of authenticity will result in meaningful outcomes that will extend far beyond their college career. Faculty, administrators, resource providers, and students should work together to maintain a fair and positive experience.

We realize that students benefit when academic integrity ground rules are established early in the course. To that end, OpenStax has created an interactive to aid with academic integrity discussions in your course.



Visit our academic integrity slider (https://view.genial.ly/61e08a7af6db870d591078c1/interactive-image-defining-academic-integrity-interactive-slider). Click and drag icons along the continuum to align these practices with your institution and course policies. You may then include the graphic on your syllabus, present it in your first course meeting, or create a handout for students. (attribution: Copyright Rice University, OpenStax, under CC BY 4.0 license)

At OpenStax we are also developing resources supporting authentic learning experiences and assessment. Please visit this book's page for updates. For an in-depth review of academic integrity strategies, we highly recommend visiting the International Center of Academic Integrity (ICAI) website at https://academicintegrity.org/ (https://openstax.org/r/academicinte).

Community Hubs

OpenStax partners with the Institute for the Study of Knowledge Management in Education (ISKME) to offer Community Hubs on OER Commons—a platform for instructors to share community-created resources that support OpenStax books, free of charge. Through our Community Hubs, instructors can upload their own materials or download resources to use in their own courses, including additional ancillaries, teaching material, multimedia, and relevant course content. We encourage instructors to join the hubs for the subjects most relevant to your teaching and research as an opportunity both to enrich your courses and to engage with other faculty. To reach the Community Hubs, visit www.oercommons.org/hubs/openstax.

Technology Partners

As allies in making high-quality learning materials accessible, our technology partners offer optional low-cost tools that are integrated with OpenStax books. To access the technology options for your text, visit your book page on OpenStax.org.

1 What Are Data and Data Science?

Figure 1.1 Petroglyphs are one of the earliest types of data generated by humanity, providing vital information about the daily life of the people who created them. (credit: modification of work "Indian petroglyphs (~100 B.C. to ~1540 A.D.) (Newspaper Rock, southeastern Utah, USA) 24" by James St. John/Flickr, CC BY 2.0)

Chapter Outline

Introduction

Many of us use the terms "data" and "data science," but not necessarily with a lot of precision. This chapter will define data science terminology and apply the terms in multiple fields. The chapter will also briefly introduce the types of technology (such as statistical software, spreadsheets, and programming languages) that data scientists use to perform their work and will then take a deeper dive into the use of Python for data analysis. The chapter should help you build a technical foundation so that you can practice the more advanced data science concepts covered in future chapters.

1.1 What Is Data Science?

Learning Outcomes

By the end of this section, you should be able to:

- 1.1.1 Describe the goals of data science.
- 1.1.2 Explain the data science cycle and goals of each step in the cycle.
- 1.1.3 Explain the role of data management in the data science process.

Data science is a field of study that investigates how to collect, manage, and analyze data of all types in order to retrieve meaningful information. Although we will describe data in more detail in Data and Datasets, you can consider *data* to be any pieces of evidence or observations that can be analyzed to provide some insights.

In its earliest days, the work of data science was spread across multiple disciplines, including statistics, mathematics, computer science, and social science. It was commonly believed that the job of data collection, management, and analysis would be carried out by different types of experts, with each job independent of one another. To be more specific, *data collection* was considered to be the province of so-called domain experts (e.g., doctors for medical data, psychologists for psychological data, business analysts for sales, logistic, and marketing data, etc.) as they had a full context of the data; *data management* was for computer scientists/engineers as they knew how to store and process data in computing systems (e.g., a single computer, a server, a data warehouse); and *data analysis* was for statisticians and mathematicians as they knew how to derive some meaningful insights from data. Technological advancement brought about the proliferation of data, muddying the boundaries between these jobs, as shown in Figure 1.2. Now, it is expected that a data scientist or data science team will have some expertise in all three domains.

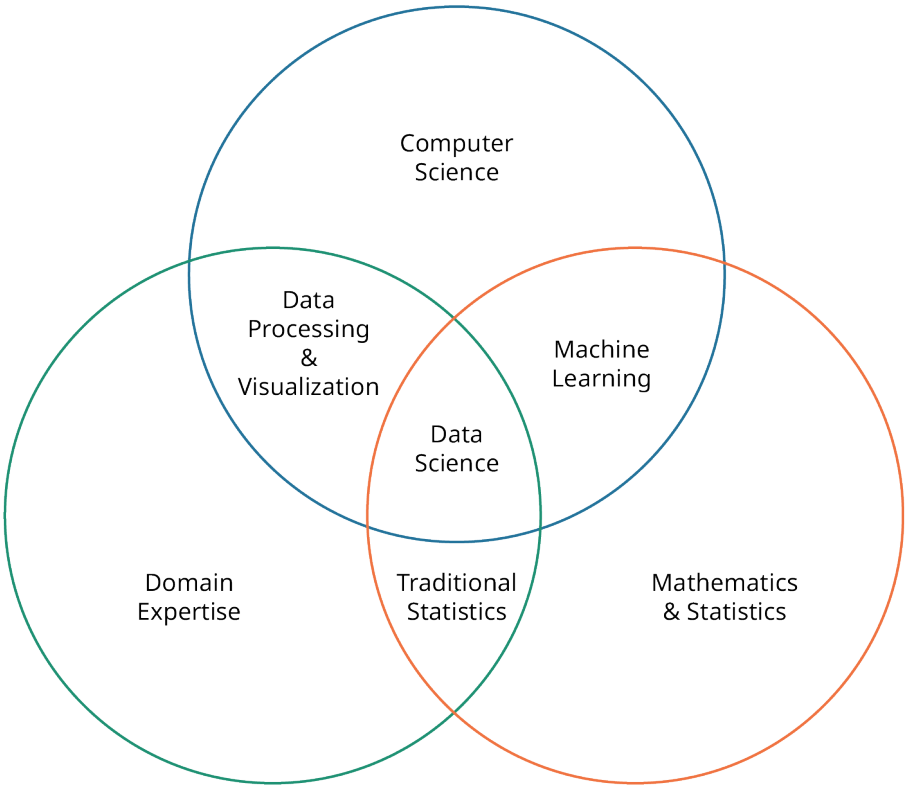


Figure 1.2 The Field of Data Science

One good example of this is the development of personal cell phones. In the past, households typically had only one landline telephone, and the only data that was generated with the telephone was the list of phone numbers called by the household members. Today the majority of consumers own a smartphone, which contains a tremendous amount of data: photos, social media contacts, videos, locations (usually), and perhaps health data (with the consumers' consent), among many other things.

Is the data from a smartphone solely collected by domain experts who are specialized in photos, videos, and such? Probably not. They are automatically logged and collected by the smartphone system itself, which is designed by computer scientists/engineers. For a health care scientist to collect data from many individuals in the "traditional" way, bringing patients into a laboratory and taking vital signs regularly over a period of time takes a lot of time and effort. A smartphone application is a more efficient and productive method, from a data collection perspective.

Data science tasks are often described as a process, and this section provides an overview for each step of that process.