Principles of

Data Science

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

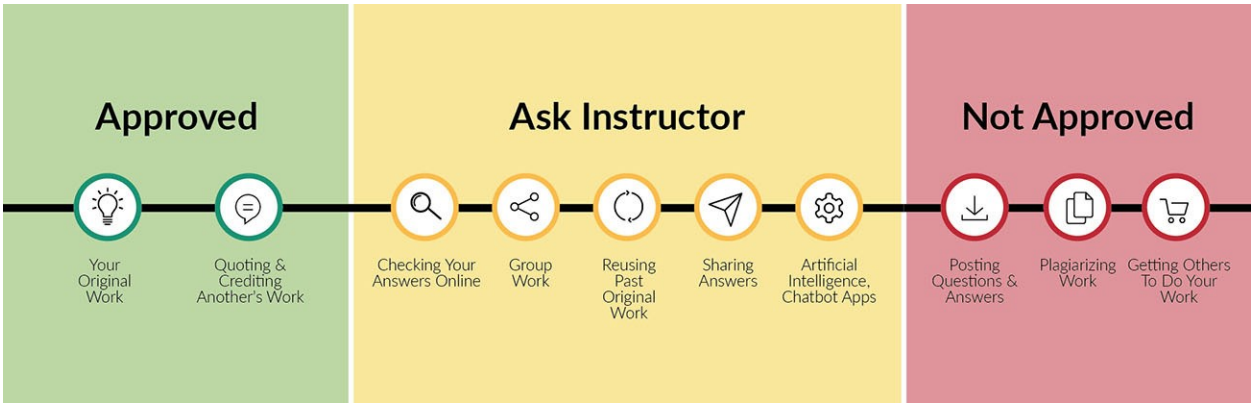
Student and Instructor Resources

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)

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 datato be any pieces of evidence or observations that can be analyzed to provide some insights.