

# Summary of *Veridical Data Science*, Chapters 1–4 \*

Shizhe Zhang (3041882158)

shizhe\_zhang@berkeley.edu

January 27, 2026

## Chapter 1: An Introduction to Veridical Data Science

Veridical data science emphasizes making data-driven decisions about the real world. Although we intensively use algorithms for data analysis these days, the core is still human judgment. Because data and models are flawed representations of reality, results must be treated as evidence rather than truth. All models are wrong, but some are useful.

This fact forces us to use our critical thinking skills and domain knowledge when facing data. We should always question the assumptions, data, conclusions. We should utilize the PCS framework, predictability, computability, and stability. These three principles help us to understand what is veridical data science and how to implement it in practice.

## Chapter 2: The Data Science Life Cycle

The data science life cycle describes the typical stages of a data project, from data collection to final results. This cycle is iterative, meaning that we should often revisit earlier stages to learn more about the data and domain. Early stages have crucial impacts on later analysis. In turn, later analysis provides feedback to earlier steps.

## Chapter 3: Setting Up Your Data Science Project

Picking the right tools is important for the analysis. There is no such one-size-fits-all tool. They all have pros and cons, and the choice depends on the specific context.

From the start of the project, we should keep documentation, reproducibility, and modular code in mind. The codes should reproduce the same results when rerun, and the documentation should explain the reasoning behind each step.

---

\*<https://vdsbook.com/>

## Chapter 4: Data Preparation

Data preparation is an important part of data science. Cleaning, filtering, transforming, and merging data all involve choices that can affect results. We need to know how the data were collected and what each feature stands for. For different problems like missing data, outliers, we need to apply domain knowledge to make reasonable decisions. We should also test whether the final conclusions are stable to these reasonable alternative preparation choices.