# Data Analysis and Machine Learning for Real-World Decision Making

STAT 214

Shizhe Zhang

shizhe_zhang@berkeley.edu

Last updated: January 22, 2026

## Contents

## ❖ Lecture 1

### 1.1 Information

Instructor: - Bin Yu (binyu@berkeley.edu)

Tu/Th 11:00am-12:30pm 20 Social Sciences Building

Office hours: Thursday 1-2pm 317 Evans

**Grading:**

- 45% lab assignments
  Lab 1: Single-person project (20%)
  Lab 2: Team project (25%)
- 10% reading assignments and selected problems from VDS book
- 2.5% peer lab review performance
- 2.5% class participation
- 5% paper presentations

- 35% final project (team project)

**GSIs:**

1. Zach Rewolinski zachrewolinski@berkeley.edu
2. Anqi Wang aqwang@berkeley.edu
3. Sequoia Andrade srandrade@berkeley.edu
4. Sean Richardson seanrichardson@berkeley.edu

## ❖ Lecture 2

### 2.1 What is Statistics?

**PCS**: predictability, computability, and stability

**Reproducibility & Stability**

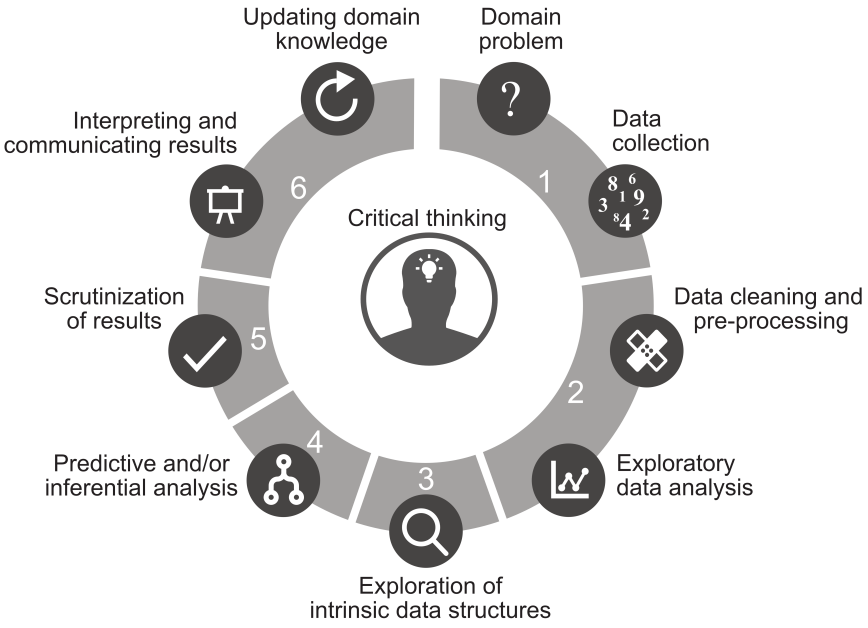


*Figure 1: data science life cycle (DSLC)*

**Definition** ***Dimension***

The dimension of the data refers to the number of variables (columns) that it contains (and sometimes also the number of rows that it contains).
So "high-dimensional data" typically refers to data that has a lot of variables

**Definition** *Rectangular Data or Tabular Data*

Data that can be represented in a spreadsheet-like. The data is arranged into columns (features/variables) and rows (observational units).

**Definition** *Reproducibility*

There is no clear definition of what it means for a data-driven result to be reproducible. Instead, there is a spectrum of definitions that range from the weakest (demonstrating that the results reemerge when the original code is rerun on the same computer) to the strongest (demonstrating that the results reemerge when a completely independent group of data scientists collect their own data and write their own code to answer the same question).
Every form of reproducibility can be viewed as a type of stability assessment (to the data collected, to the code written, to the person who conducted the analysis, etc.) and/or predictability assessment (if re-evaluation involves showing that the results reemerge using new data). You are encouraged to demonstrate the strongest form of reproducibility for which you have the resources (even if this is just demonstrating that your results reemerge when your code is rerun in a fresh R session).

ABC abc

**Textbooks**

1. "Veridical data science: the practice of responsible data analysis and decision-making" by Bin Yu and Rebecca Barter (MIT Press, in-press) (free online version at vdsbook.com) (required)

2. Statistical models, David Freedman (Cambridge Press, 2009, 2nd Ed.) (required). (open-source pdf)

3. The elements of statistical learning, Trevor Hastie, Rob Tibshirani, Jerome Friedman (Springer, 2016, 2nd Ed.) (recommended). (open-source pdf)