Data 100: Principles and Techniques of Data Science

Michael Pham

Spring 2024

Contents

Week 1
Introduction to Data Science

The purpose of computing is insight, not numbers.

— R. Hamming

1.1 Lecture 1 – 01/16/24

The course website is located at: https://ds100.org/sp24/.

1.1.1 Course Overview

Why Data Science Matters

Data is used everywhere, from science to sports to medicine. Claims using data also comes up often within discussions (especially about important issues).

Furthermore, Data Science enhances critical thinking. The world is complicated, and decisions are hard. This field fundamentally facilitates decision-making by quantitatively balancing trade-offs.

In order to quantify things reliably, we have to:

- Find relevant data;
- Recognize the limitations of said data;
- Ask the right questions;
- Make reasonable assumptions;
- Conduct appropriate analysis; and
- Synthesize and explain our insights.

At each step of this process, we must apply critical thinking and consider how our decisions can affect others.

What is Data Science?

Definition 1.1 (Data Science). Data Science is the application of data-centric, computational, and inferential thinking to:

- Under the world (science), and
- Solve problems (engineering).

We note that good data analysis is **not**:

- Simple applications of a statistics recipe.
- Simple application of software.

There are many tools out there for data science, but they are ultimately just tools; we are the ones doing the important thinking.

1.1.2 Course Outline

Prerequisites

The official prerequisites are:

- Data 8;
- CS 61A, Data C88C, or Engin 7; and
- EE 16A, Math 54, or Stat 89A.

Topics (Tentative)

The tentative list of topics that will be covered in this course is:

Tentative Topics

- Pandas and NumPy
- Relational Databases and SQL
- Exploratory Data Analysis
- Regular Expressions
- Visualization
  - matplotlib
  - Seaborn
  - plotly
- Sampling
- Probability and random variables
- Model design and loss formulation
- Linear Regression
- Feature Engineering
- Regularization, Bias-Variance Tradeoff, and Cross-Validation
- Gradient Descent
- Data Science in the Physical World
- Logistic Regression
- Clustering
- PCA

Course Components

With respect to lectures and assignments, the course is structured as follows:

Course Components

Mo	Tu	We	Th	Fr
	Live Lecture		Live Lecture	
	Discussion	Discussion		
	Office Hours	Office Hours	Office Hours	Office Hours
			Homework N-1 due	Homework N released
	Lab N-1 due			Lab N released

For lectures, note that there attendance is mandatory; participation will be graded on a 0/1 basis:

- Synchronous Participation: complete at least one participation poll question during the live lecture timeslot (11:00am-12:30pm, Tuesdays and Thursdays). As long as you submit a response to at least one poll question in this timeframe, you will receive synchronous attendance credit.
- Asynchronous Participation: complete all participation poll questions from the link provided on the course website within one week of the corresponding lecture.
- In both cases, participation is graded on completion, not correctness.

Also, if we submit all participation polls over the semester, there will be a 0.5% bonus points applied to the final overall grade.

Grading

The grading scheme for this class is as follows:

Grading Scheme

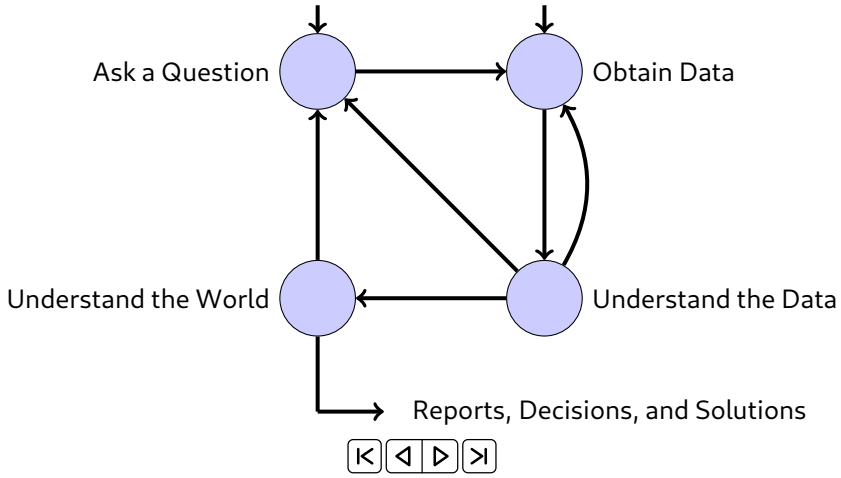
Category	
Homeworks	25%
Projects	10%
Labs	5%
Discussions	-
Lecture Participation	5%
Midterm Exam	22.5%
Final Exam	32.5%

Important:

- Midterm: Thursday, March 7, 7-9 PM PST.
- Final: Thursday, May 9, 8-11 AM PST.

1.1.3 The Data Science Lifecycle

The data science lifecycle is a high-level description of the data science workflow. Note in the diagram below that there are two distinct entry points.



The Data Science Lifecycle goes as follows:

1. Question/Problem Formulation:
    - What do we want to know?
    - What problems are we trying to solve?
    - What hypotheses do we want to test?
    - What are our metrics for success?
2. Data Acquisition and Cleaning
    - What data do we have and what data do we need?
    - How will we sample more data?
    - Is our data representative of the population we want to study?
3. Exploratory Data Analysis and Visualization
    - How is our data organized, and what does it contain?
    - Do we already have the relevant data?
    - What are the biases, anomalies, or other issues with the data?
    - How do we transform the data to enable effective analysis?
4. Prediction and Inference
    - What does the data say about the world?
    - Does it answer our questions or accurately solve the problem?
    - How robust are our conclusions and can we trust the predictions?