Welcome to STAT 4710J

Data Science and Analytics with Python

Instructor: Ailin Zhang (ailin.zhang@sjtu.edu.cn)

Lecture 1: Course Overview

Agenda

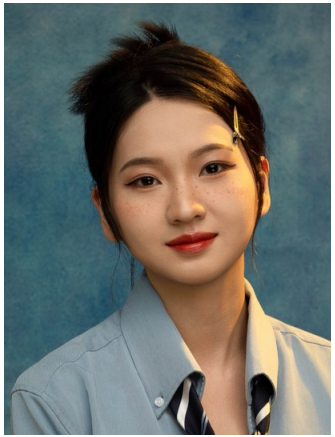
- Meet your Instructor and TAs
- What is Data Science?
- Goals for the course
- Course Logistics
- Meet your classmates

Meet your Instructor

Ailin Zhang

- Background: PhD in Geophysics from UCLA (2019). Formerly data scientist @ ExxonMobil
- Research Interests: Data-driven solutions to geoscience problems: earthquake rupture, oil and gas exploration, seismic signal processing.
- OH: By Appointment @ Longbin Building 437B
- I am teaching the following courses in the semester
    - STAT 4710
    - STAT 4130: Applied Regression Analysis

Meet your TAs



- **Name：** Li Li
- **Major/Year：** ECE Senior
- **Minor：** Data Science
- **Email：** lily-l@sjtu.edu.cn
- **Future Direction：** Information Networking

Related Course Studied：

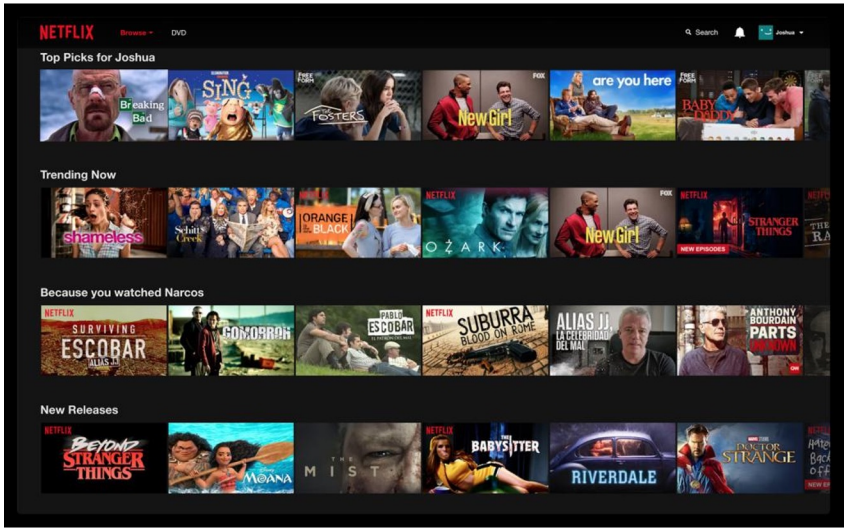
- STAT4710J
- STAT4130J
- STAT4060J
- STAT3060J

Meet your TAs

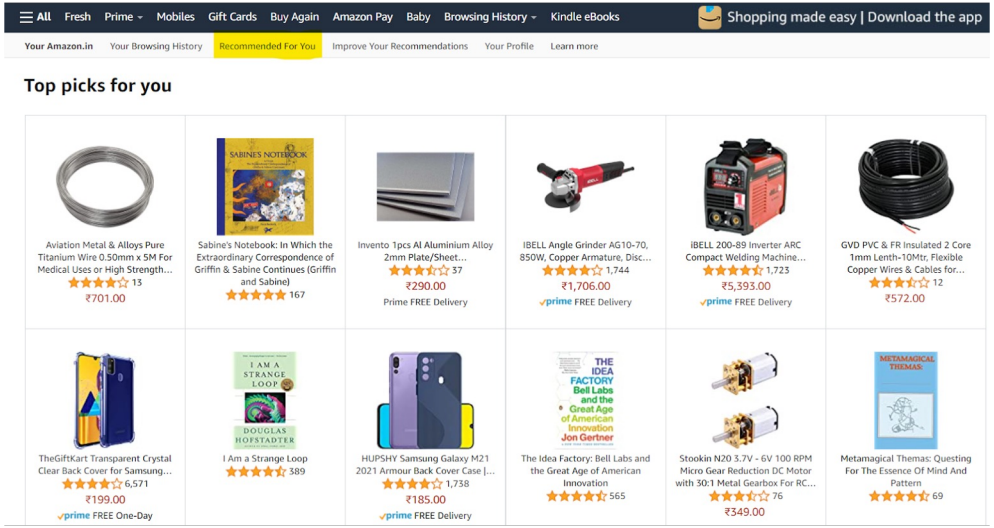
- **Name**: Wang, Peiran (王沛然)
- **Major/Year**: ECE+DS Senior
- **Email**: phoenix_wang@sjtu.edu.cn
- **Future Direction**: Statistics
- **Related Course Studied**:
  - ECE4710J /STATS4710J
  - STATS413 Applied Regression Analysis
  - STATS415 Statistical Learning

First time as a TA. Hope to make progress with the class together

Some Examples of Data Science



amazon



Some Examples of Data Science

Data science enhances critical thinking

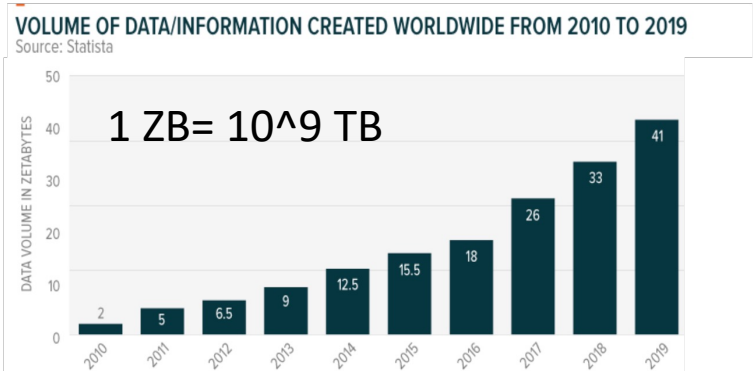
The world is complicated! Decisions are hard.

Data is used everywhere to answer hard questions and make tough decisions:

- Science
- Medicine
- Social science
- Engineering
- Sports

Technology Trends

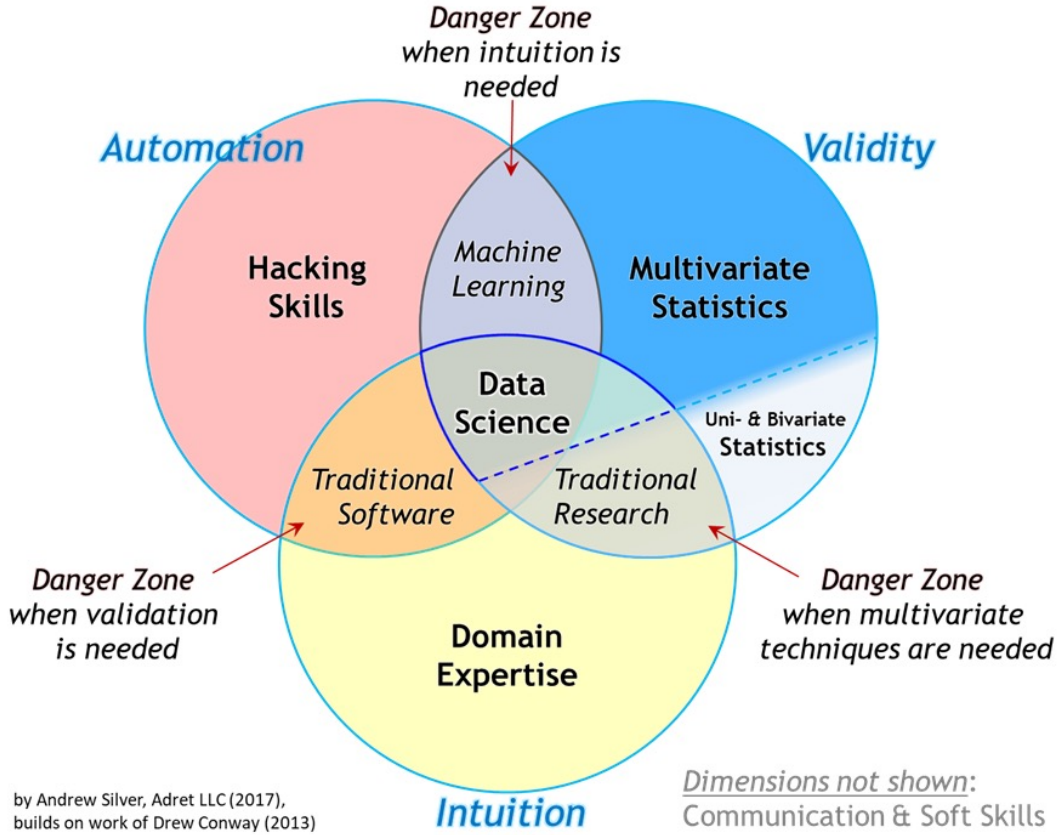
- 2020s: ?
- 2010s: Data Industry
  - Collect and sell information
- 2000s: Internet Industry
  - Online retailers and services
- 1990s: Software Industry
  - Sold computer software
- 1980s: Hardware Industry
  - Sold computers



What is Data Science? (in a few keywords)

Waiting for responses ...

What is data science?



Data Science Venn Diagram

What is Data Science?

Data science is a fundamentally interdisciplinary field

Wikipedia:

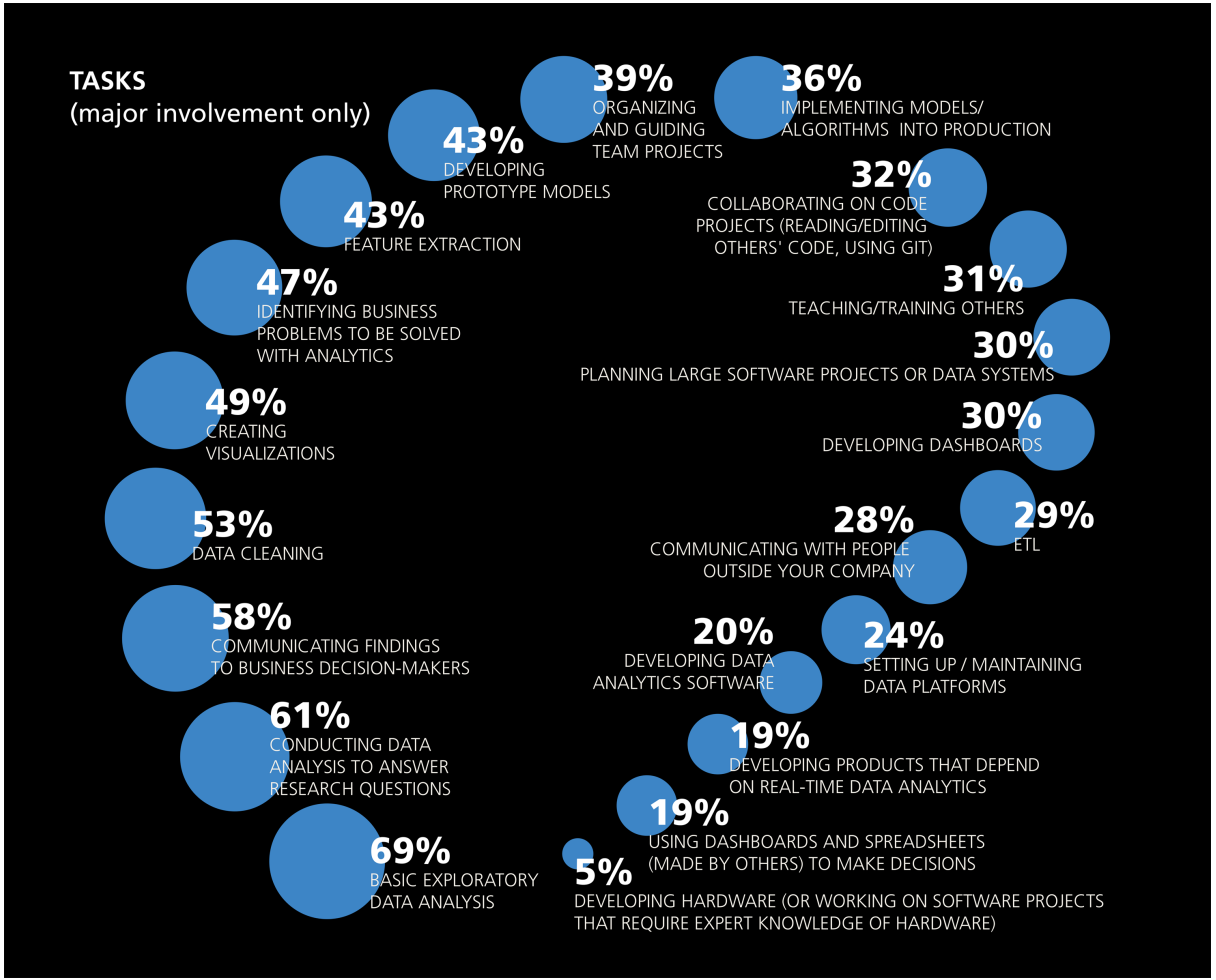
"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains."

Data Science is the application of data centric, computational, and inferential thinking to:

- Understand the world (science).
- Solve problems (engineering).

What tasks do data scientists do regularly?

https://www.oreilly.com/radar/2016-data-science-salary-survey-results/



Insight

Good data analysis is not:

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

There are many **tools** out there for data science, but they are merely tools.

- **They don't do any of the important thinking!**

"The purpose of computing is insight, not numbers." - R. Hamming. *Numerical Methods for Scientists and Engineers (1962).*

The Darker Side of Data Science?

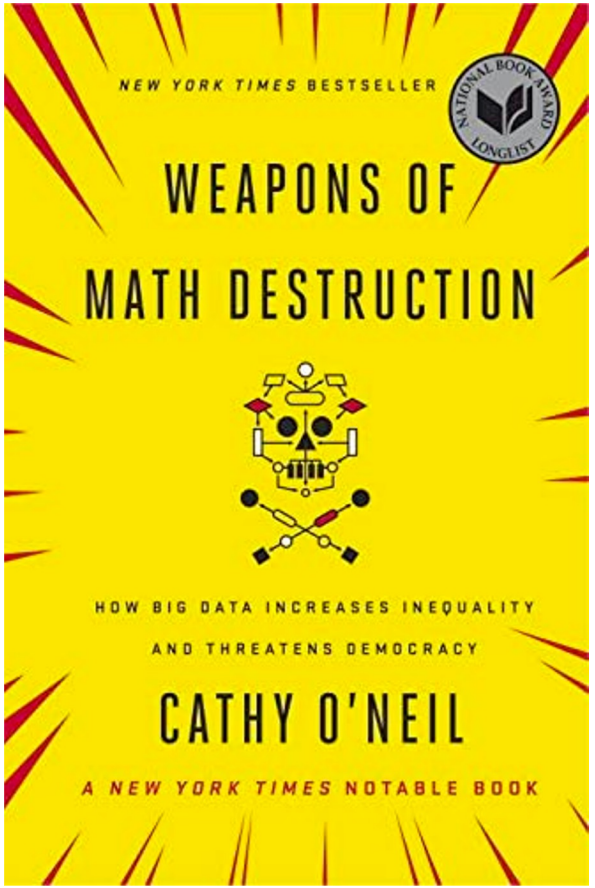
Obscuring complex decisions:

- Mortgage-backed securities → market crash
- job advancement

Reinforcing historical trends and biases:

- Hiring based on previous hiring data
- Recidivism and racially biased sentencing
- Social media, news, and politics

We will discuss the ethics of data science throughout the class!



Course goals

Prepare

Prepare students for advanced courses in **data management**, **machine learning**, and **statistics**, by providing the necessary foundation and context.

Enable

Enable students to start careers as data scientists by providing experience working with **real-world data, tools, and techniques**. Provide some help for your interviews!

Empower

Empower students to apply computational and inferential thinking to address **real-world problems** in their lives.

Tentative List of Topics to be Covered

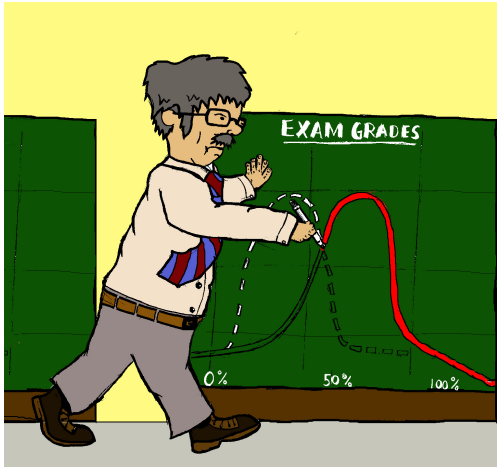
- Pandas and NumPy
- Relational Databases & SQL
- Exploratory Data Analysis
- Regular Expressions
- Visualization
  - matplotlib
  - Seaborn
  - plotly
- Sampling
- Probability and random variables
- Web scrapping and HTML
- Model design and loss formulation
- Linear Regression
- Feature Engineering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

Prerequisites

- **These are not strictly enforced**, but we will not be teaching:
  - How to use Python.
  - How to use Jupyter notebooks.
  - Basic linear algebra, probability and statistics
- We will use homework 1 to recap some key concepts as a warm-up.
- We are here to help!
  - We really want you to succeed in this class.
  - Feel free to reach out with any questions or concerns you have.
  - Office hour: TBD

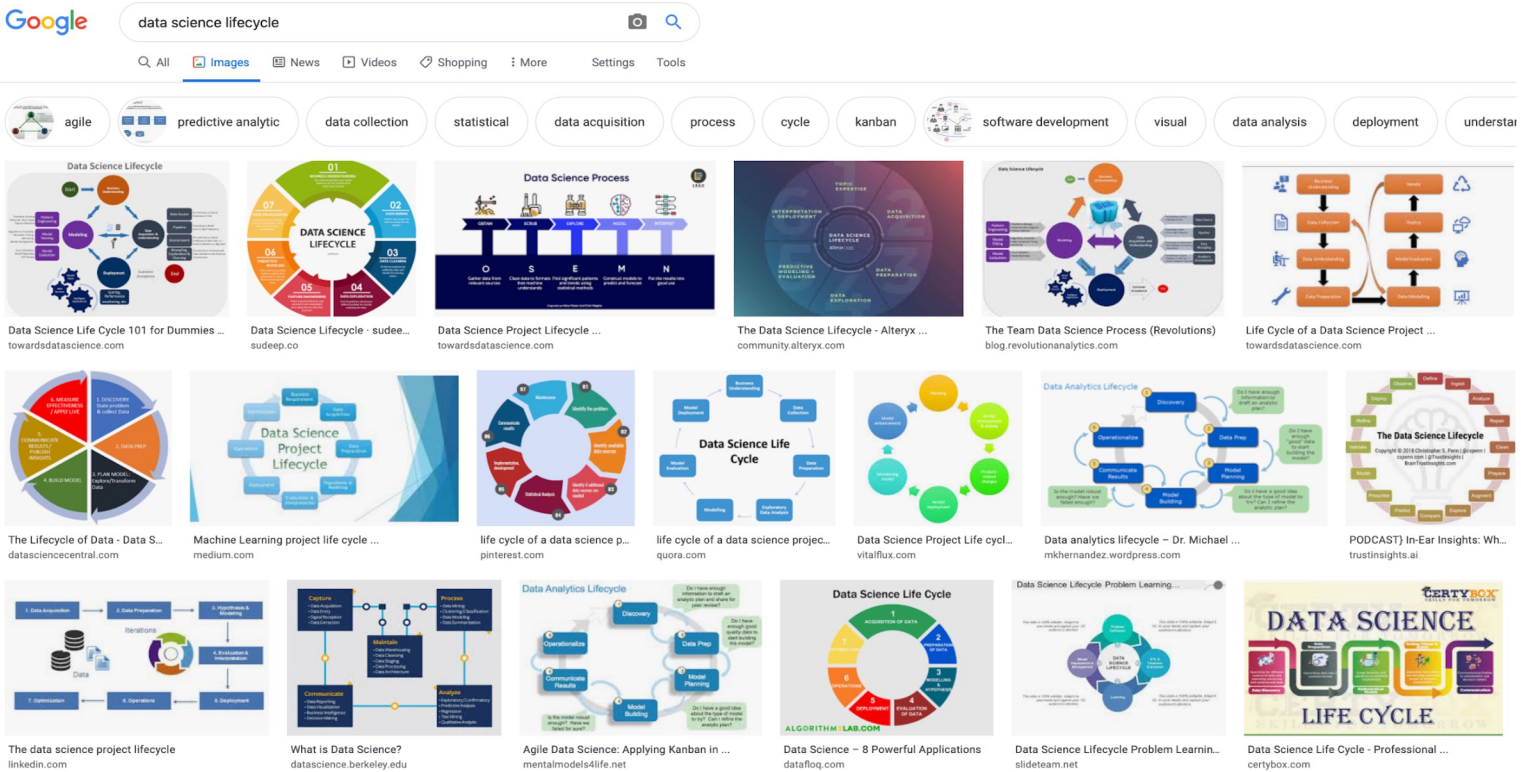
Grading

- I reserve the right to curve the scale if there are less than 30% of students with grades ≥ A.
  - **5%** Homework
  - **30%** Lab (10% participation + 20% performance)
  - **15%** Project
  - **40%** Midterm (~week 7,before July)
  - **10%** Final (Take home challenge)
  - **1%*** Extra Credit

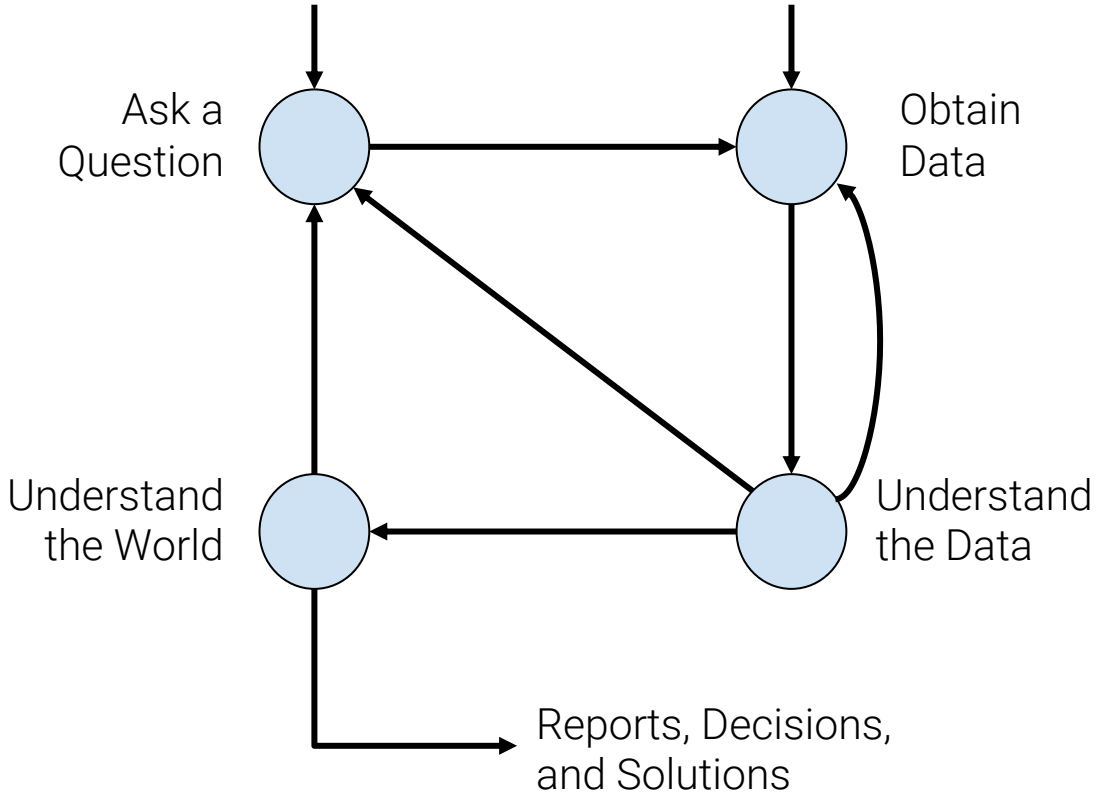


- You can use Chatgpt but you must state it in your homework/ project.

Workflow: Data Science Lifecycle



Data science lifecycle

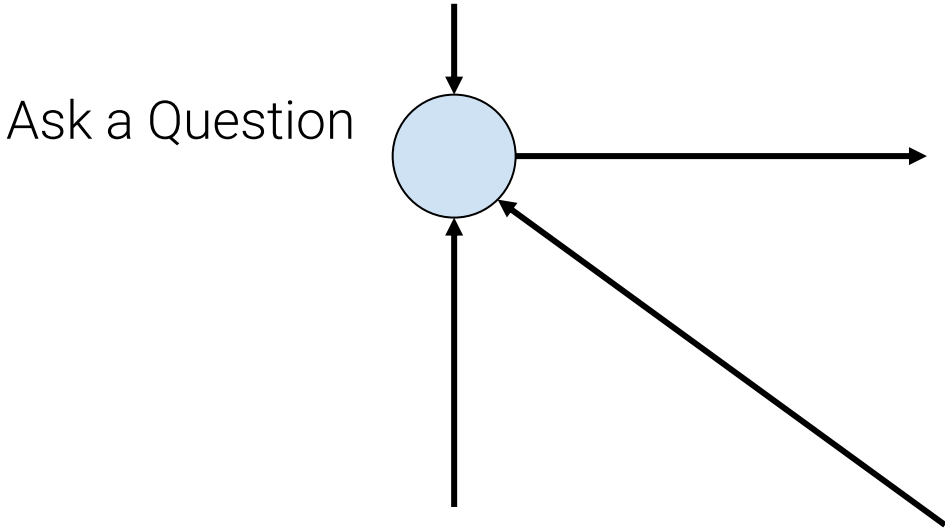


The data science lifecycle is a **high-level description** of the data science workflow.

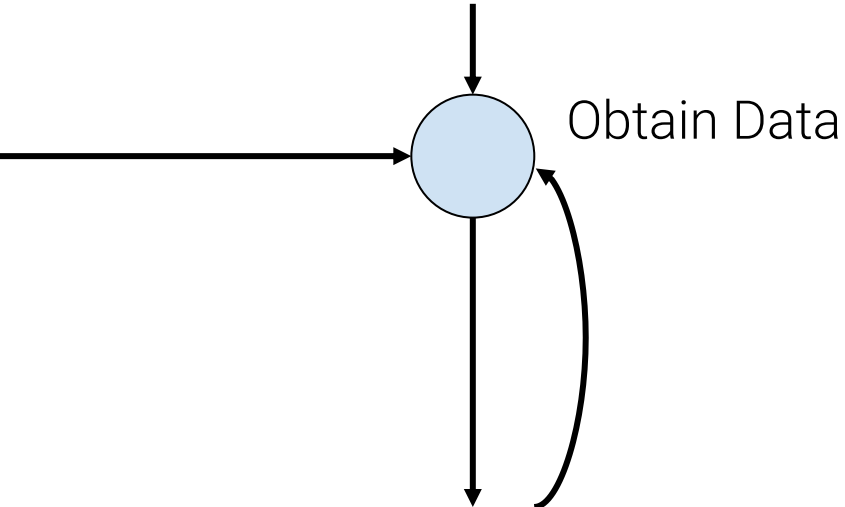
Note the two distinct entry points!

1. Question/Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics for success?

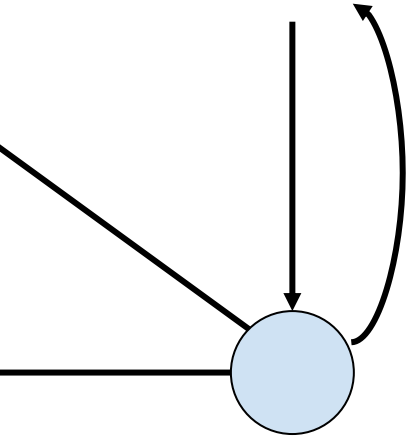


2. Data Acquisition and Cleaning



- 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 & Visualization

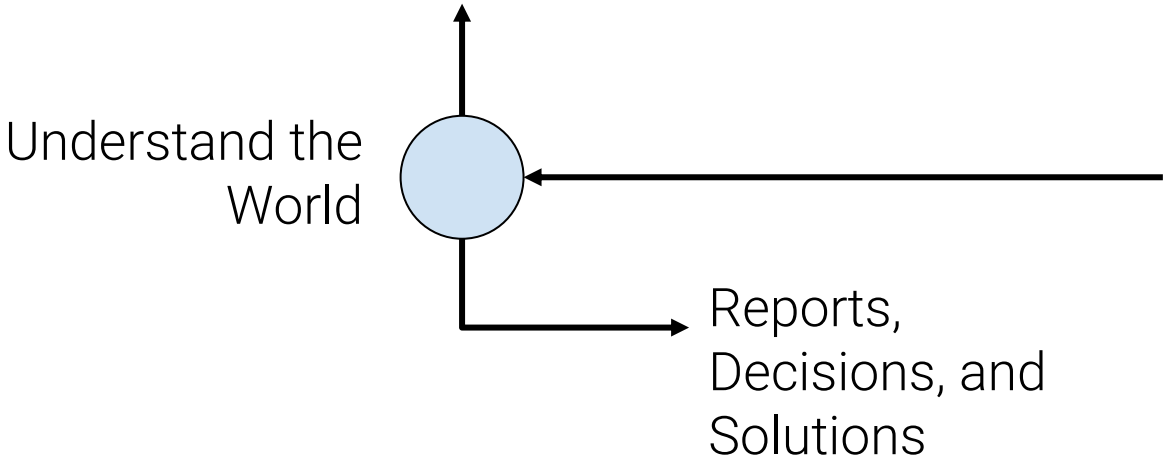


Understand the Data

- How is our data organized and what does it contain?
- Do we already have 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?



Meet You!

- Name, Year, Major
- What is your experience with Data Science (courses, projects, research etc…)?
- What do you want to get out of this class?
- Any other questions?