Welcome to ECE 4710J

Introduction to Data Science (Jiao Tong Global Classroom!)

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

Lecture 1: Course Overview

Agenda

- Meet your instructor
- What is Data Science?
- Goals for the course
- Course Logistics
- Meet your classmates

仅限团队内部成员加入

该二维码 1 年内 (2023/5/8前)有效

Meet your instructor

Ailin Zhang

- Joined JI in 2021.9
- 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.

GRPI team effectiveness model

Course goals, desired results, expectations.

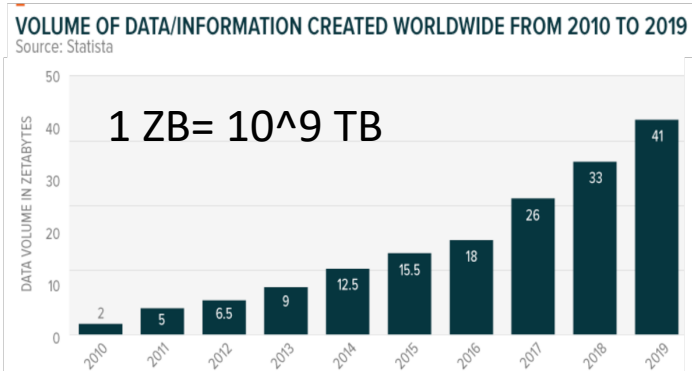
Role clarity, responsibility, and boundaries

Class norms, styles, agreement on course content and reward systems

Trust, Psychological Safety, Flexibility, Courage of conviction

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

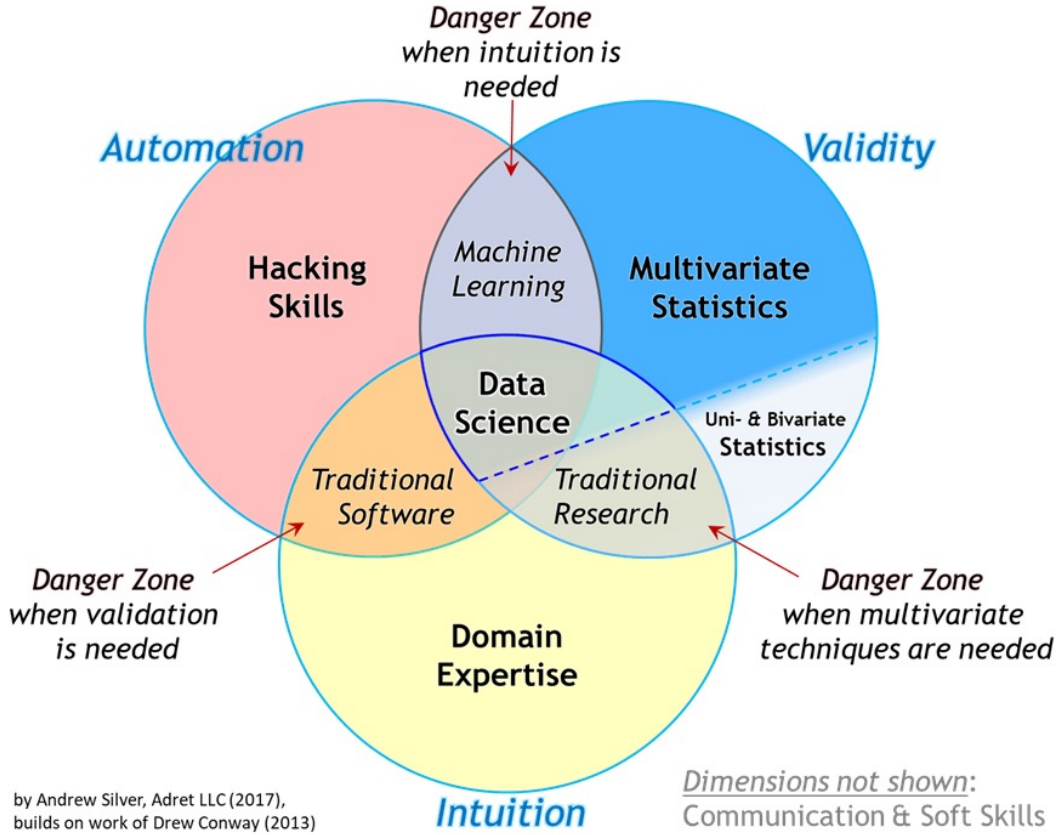


Go to **www.menti.com** and use the code **5507 6439**

What is Data Science? (in a few keywords)

Mentimeter

What is data science?



Data Science Venn Diagram

What is Data Science?

Data science is a fundamentally interdisciplinary field

Wikipedia:

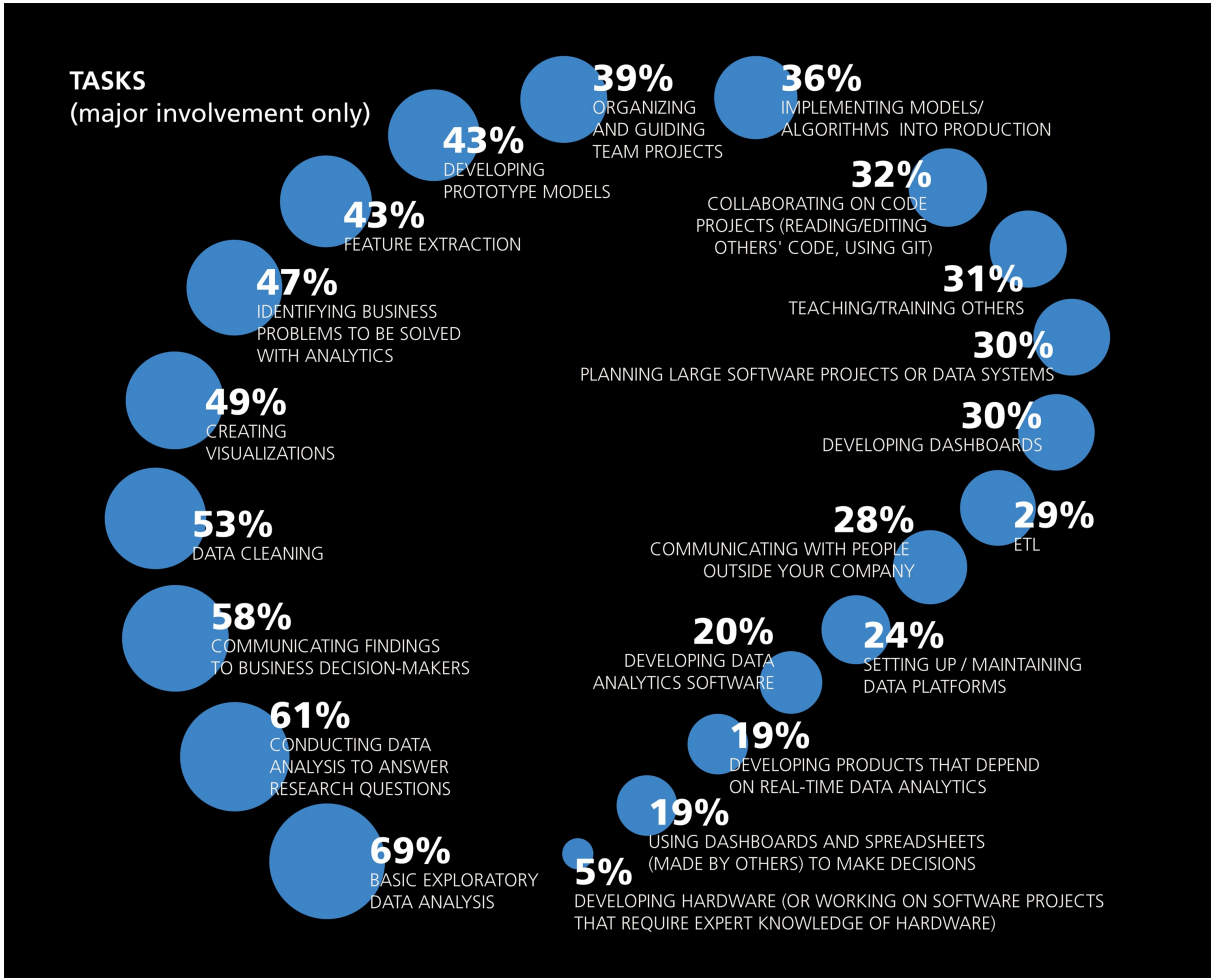
"Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains."

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).*

Example questions in data science

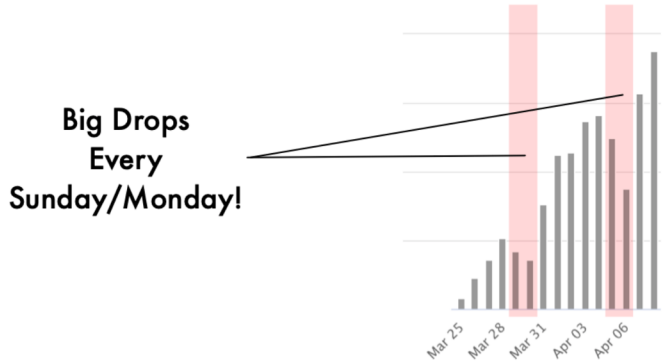
Some (broad) questions we might try to answer with data science:

- Is the world getting better or worse?
- What should we eat to avoid dying early of heart disease?
- Do immigrants from poor countries have a positive or negative impact on the economy?
- Are people vaccinated against COVID-19 also protected from new variants?

Data science drives policy and public understanding

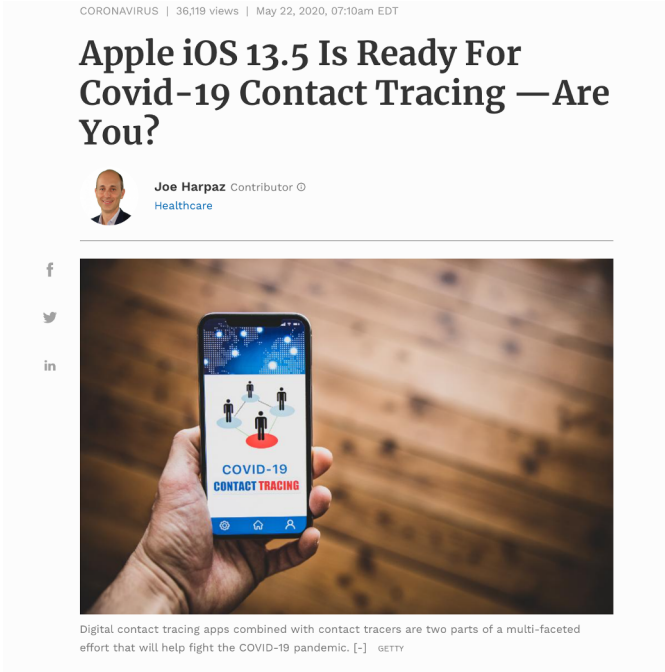
There are real-world implications of the work we do as data scientists.

Let's take a look at the daily numbers reported by the United Kingdom:



Daily Deaths due to COVID in the UK from https://www.worldometers.info/coronavirus/country/uk/

The problem is that this weekly cycle is fake. It's an artifact of how the data is collected and reported.



CORONAVIRUS | 36,119 views | May 22, 2020, 07:10am EDT

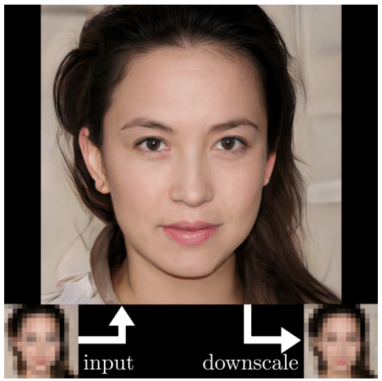
Apple iOS 13.5 Is Ready For Covid-19 Contact Tracing —Are You?

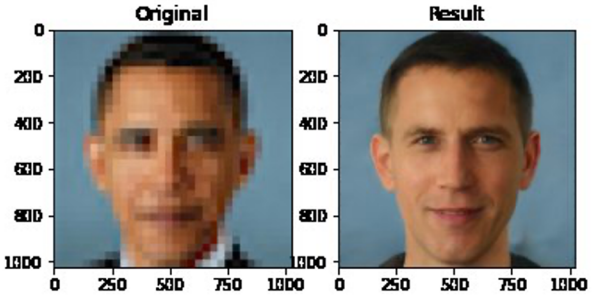
Joe Harpaz Contributor
Healthcare

Digital contact tracing apps combined with contact tracers are two parts of a multi-faceted effort that will help fight the COVID-19 pandemic. [-] GETTY

Unconscious bias is real – be mindful of it

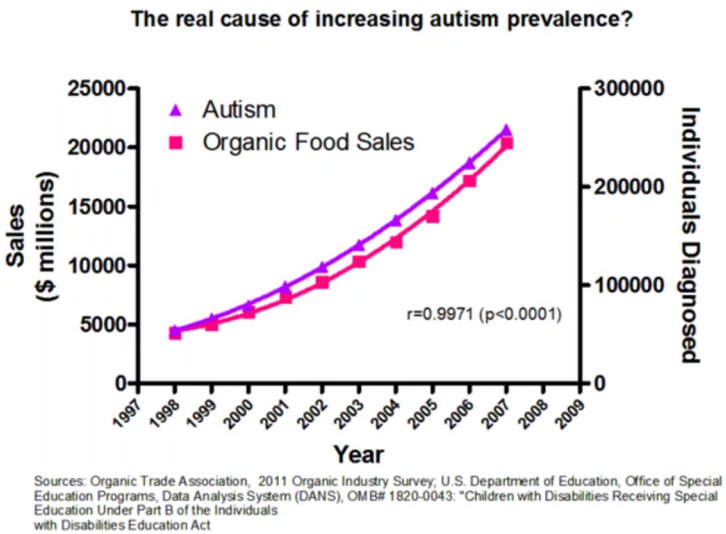
A “depixelizer” was built that takes pixelated images and generates images that are perceptually realistic and downscale correctly.





What do you notice? **Why** might this be happening?

Question what you see



Are autism rates and organic food sales inherently related? Seems unlikely.

Critical thinking is the most in-demand!

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

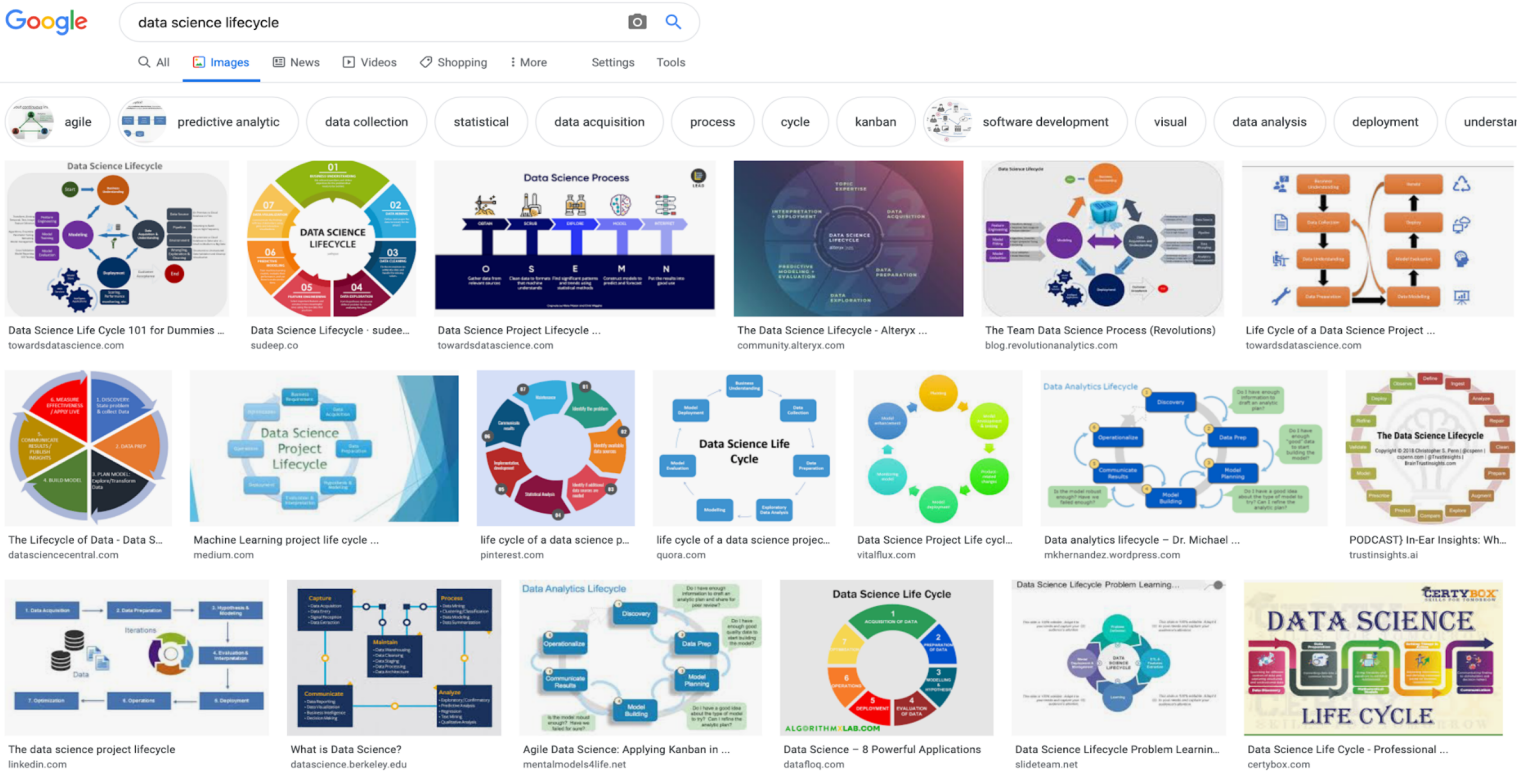
Empower students to apply computational and inferential thinking to address **real-world problems** in their lives.

Topics to be covered

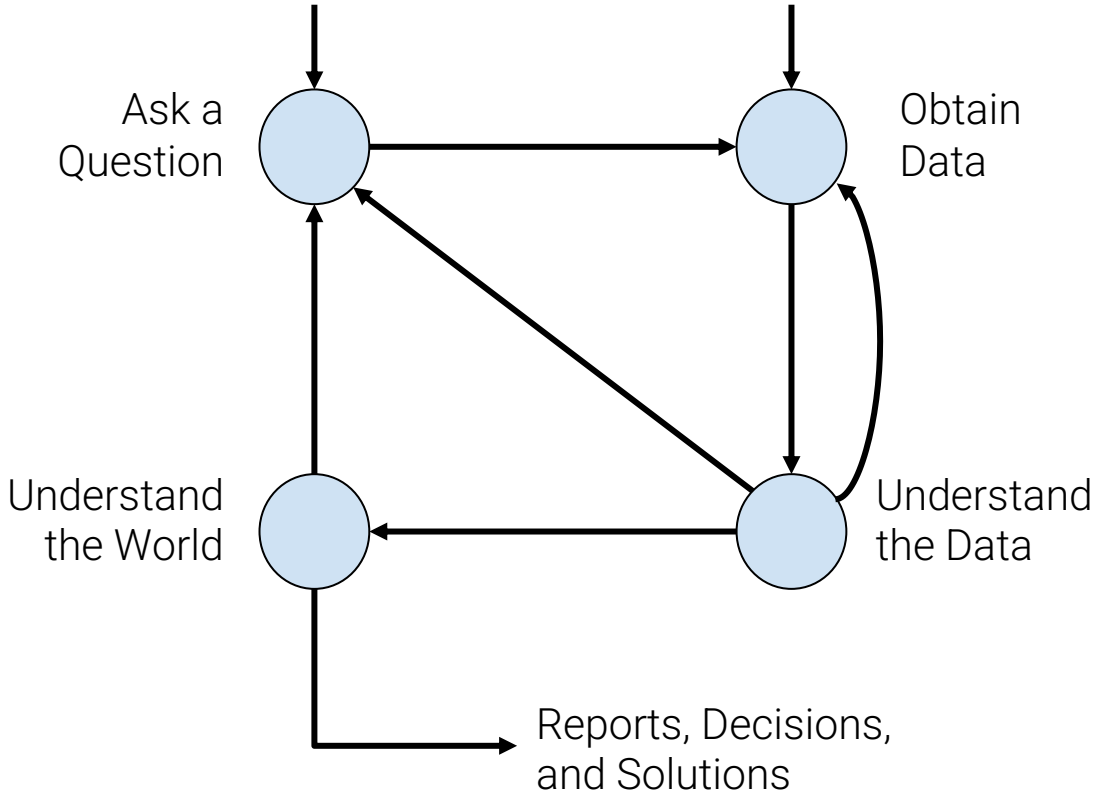
- Pandas and NumPy
- Exploratory Data Analysis
- Regular Expressions
- Visualization (Matplotlib …)
- Sampling
- Probability and random variables
- Model design and loss formulation

- Linear Regression
- Feature Engineering
- Regularization, Bias-Variance Tradeoff, Cross-Validation
- Gradient Descent
- Logistic Regression
- Decision Trees and Random Forests
- PCA
- Clustering

Workflow: Data Science Lifecycle



Data science lifecycle

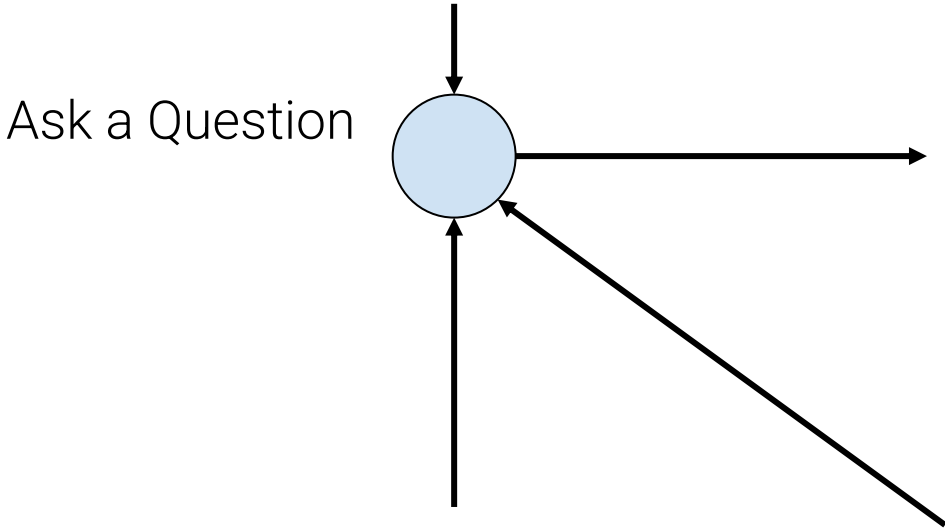


The data science lifecycle is a **high-level description** of the data science workflow.

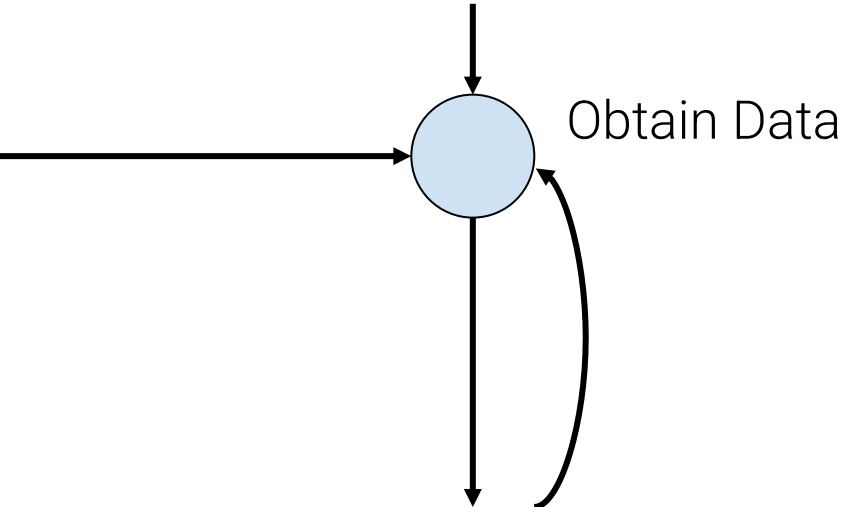
Note the two distinct entry points!

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?

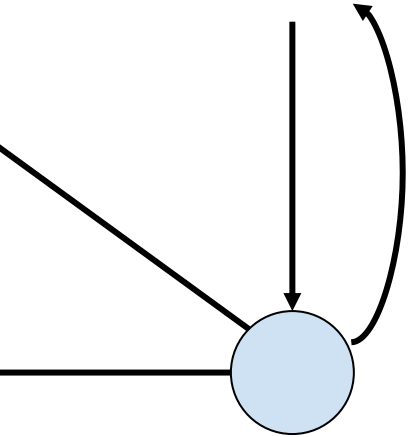


2. Data Acquisition and Cleaning



- 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

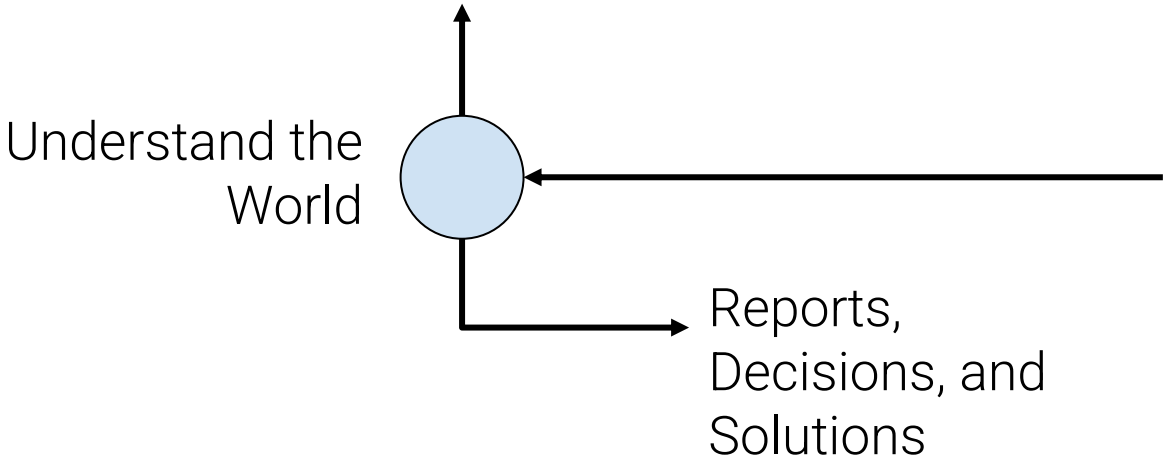


Understand the Data

- 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?

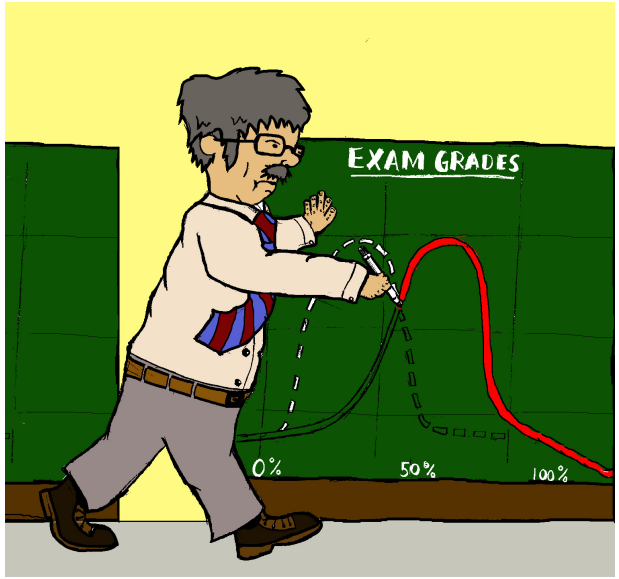


Prerequisites

- **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: Monday 3-5 PM / Feishu (24-hour policy)

Grading

- I reserve the right to curve the scale if there are less than 30% of students with grades ≥ A.
  - **25%** Homework (5-7 submissions)
  - **25%** Project
  - **30%** Midterm (Week 6 - Online)
  - **20%** Final (Week 13 – Take home)
  - **3%*** Extra Credit



Meet You!

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