CSC380: Principles of Data Science

Introduction and Course Overview

Xinchen Yu

Today’s Plan

- Data Science Introduction
- Course and syllabus overview

Course Staff

Course Instructor

- Xinchen Yu, Ph.D.
- Research:
  - Natural Language Processing, Computational Social Science
- Office hours:
  - Thursdays, 12:00pm-2:00pm (priority window 1:00-2:00pm)

Course TAs:

- Eric Duong, Saiful Salim, Tian Tan

What is “Data Science”?

My Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

What is “Data Science”?

My Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

amazon

What is “Data Science”?

My Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

amazon

Examples:

- Do people in college towns tend to buy more laptops than people in other areas?

What is “Data Science”?

My Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

amazon

Examples:

- Do people in college towns tend to buy more laptops than people in other areas?
- Find out top-10 sales categories for each age group.
- Summarize product reviews with respect to product quality.

What is “Data Science”?

My Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

amazon

Examples:

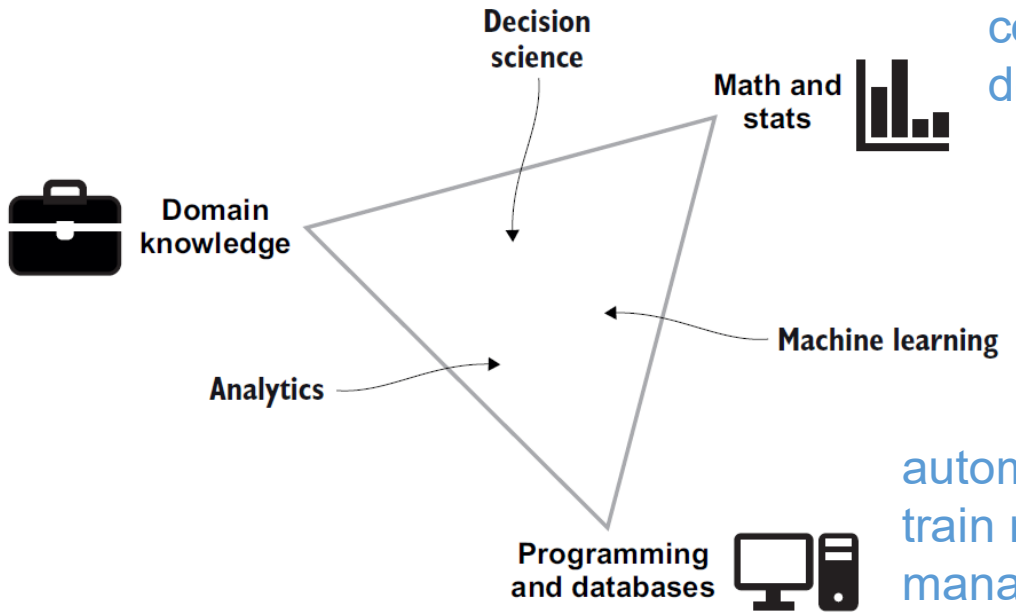
- Do people in college towns tend to buy more laptops than people in other areas?
- Find out top-10 sales categories for each age group.
- Summarize product reviews with respect to product quality.
- If we recommend pens to users from college town, how much will it increase our revenue?

What is “Data Science”?

My Definition: *The process of using data to (1) answer questions, (2) extract knowledge, and (3) predict future outcomes.*

amazon:

how customers behave



compute average/variance
design statistical models

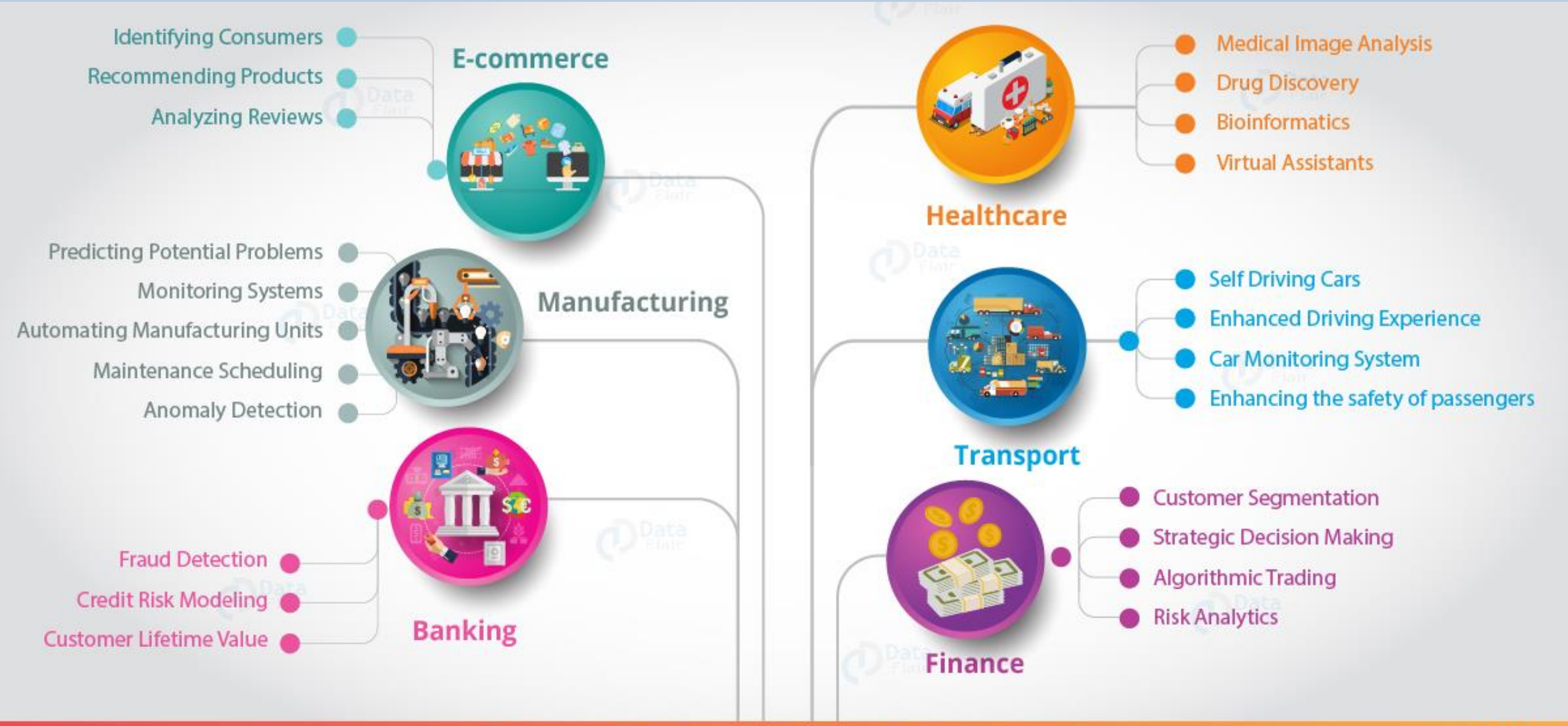
automate analyses
train models
manage data (store, retrieve)

[Source: Robinson, E. and Nolis, J.]

Data Science Is:

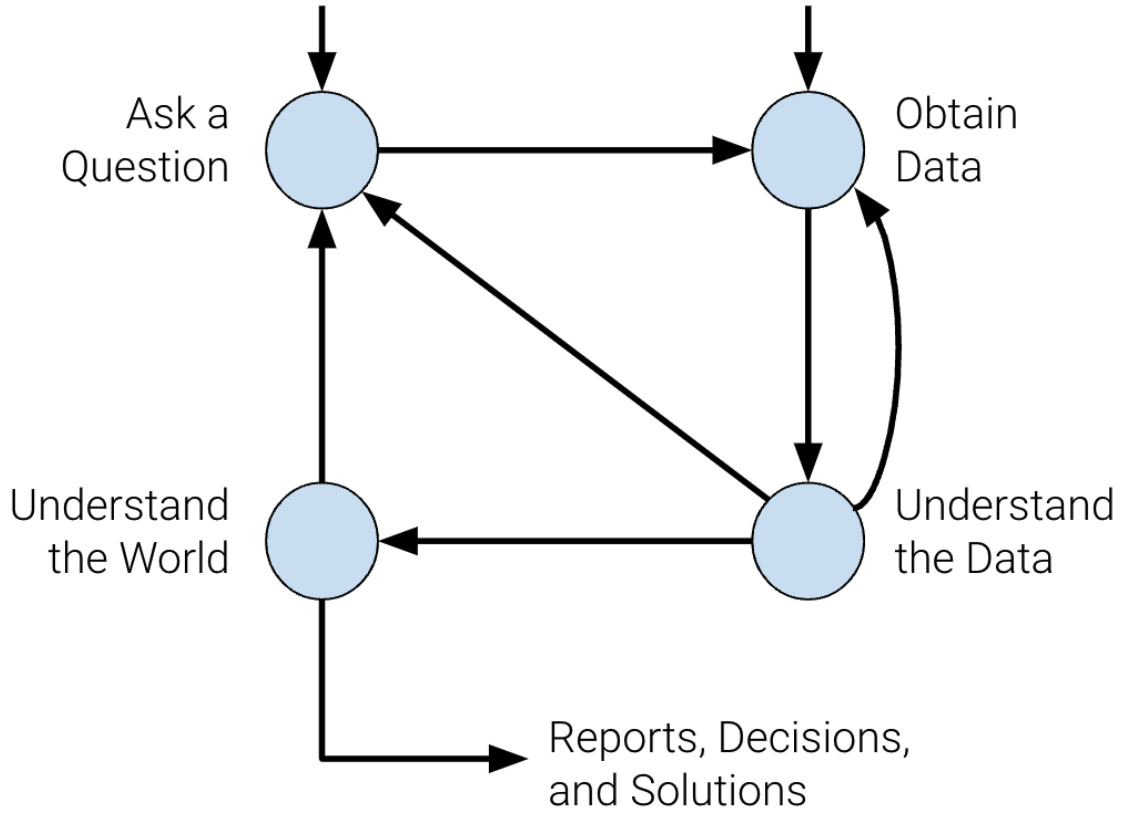
- **Interdisciplinary**: Combines tools and techniques from Math / Statistics / CS
- **Exploratory**: Understanding data requires creative exploration and visualization
- **Applied Statistics & Probability** + extra stuff to handle, process, and visualize data

What is “Data Science”?



Data Science Applications

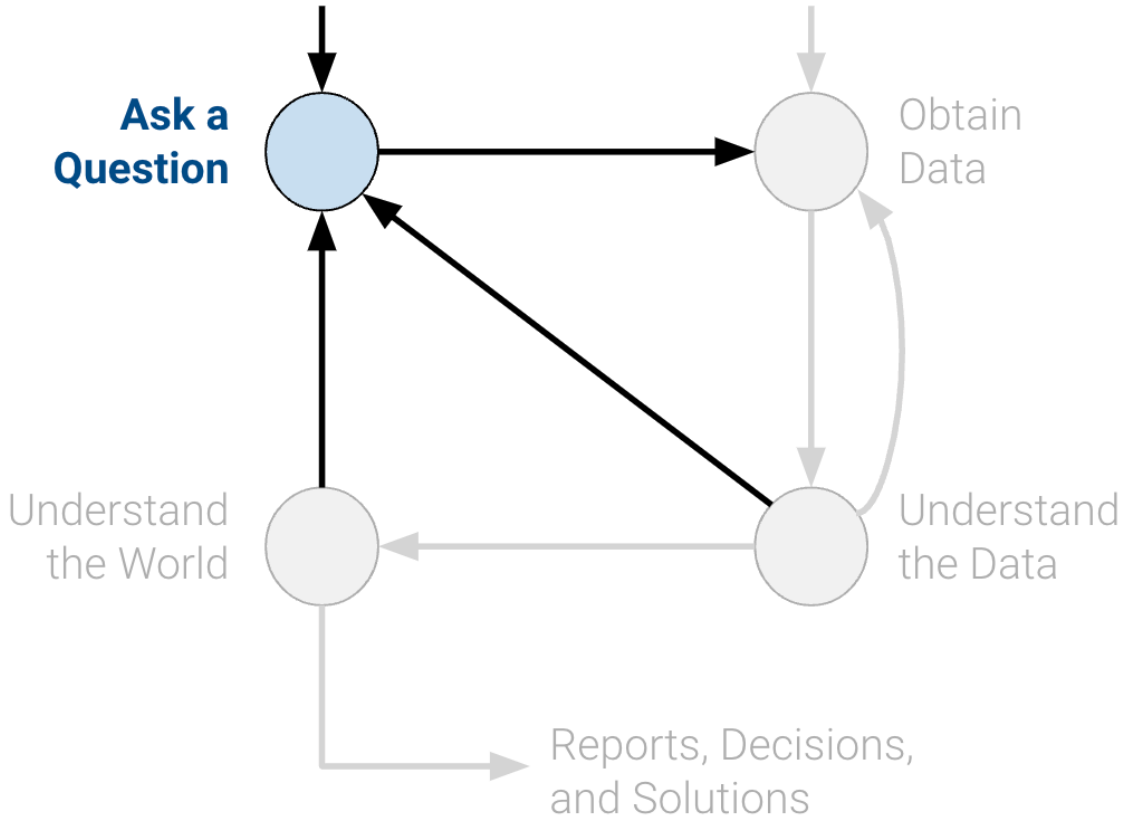
The data science life cycle



Data 100, UC Berkeley

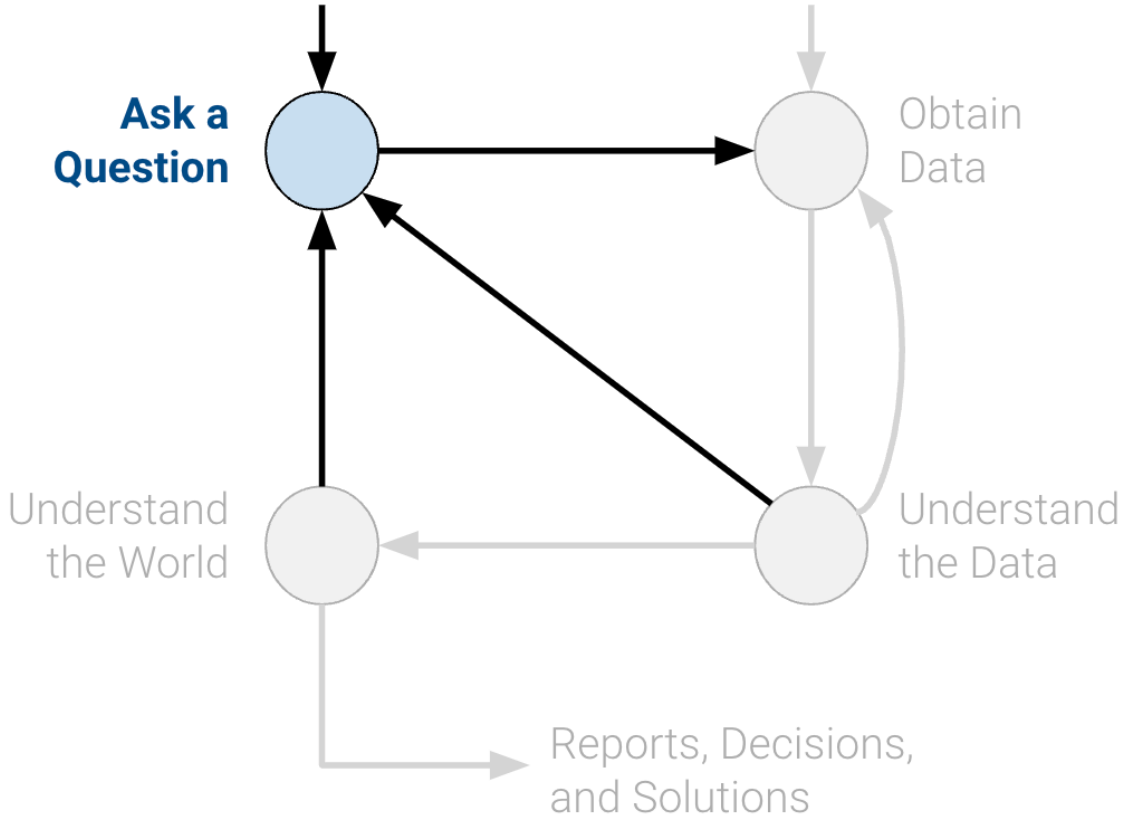
The data science life cycle

- What do we want to know?
- What problems are we solving?
- What hypotheses are we testing?
- What are our success metrics?



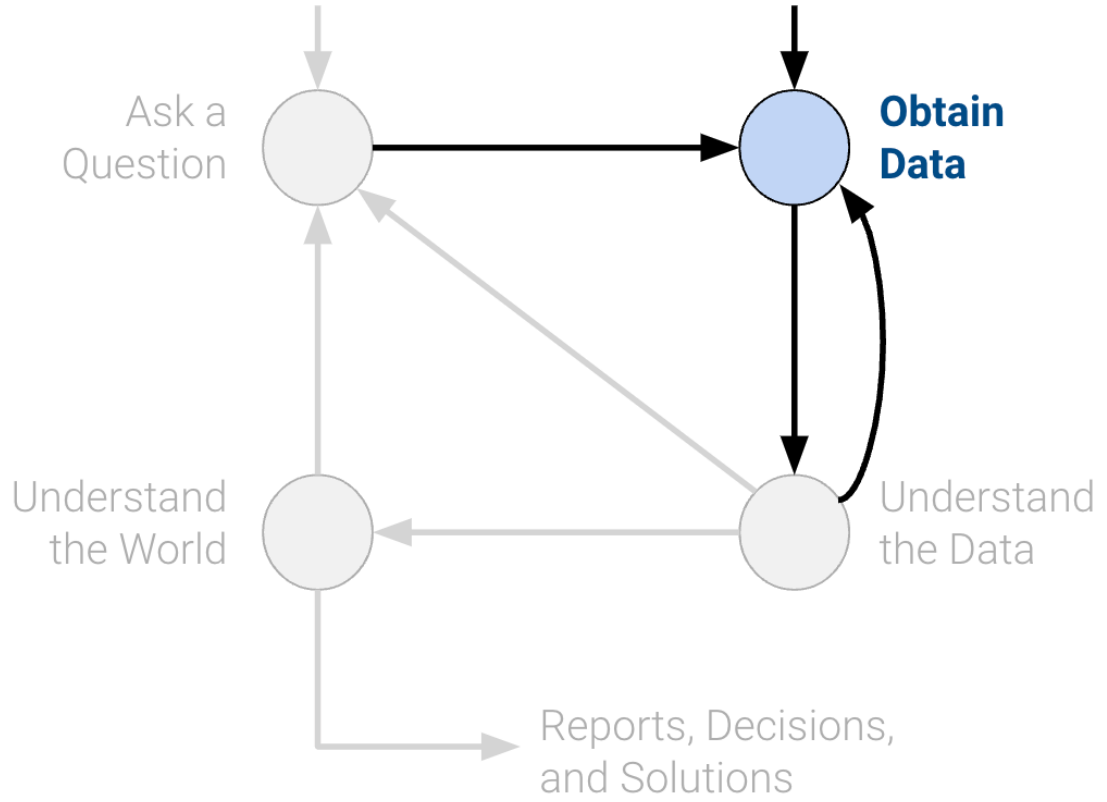
An example

- What do we want to know?
  - What is the age distribution of lead actors/actresses in movies with ratings above 8.5?
  - What does the collaboration (co-cast) network of actors and actresses look like?



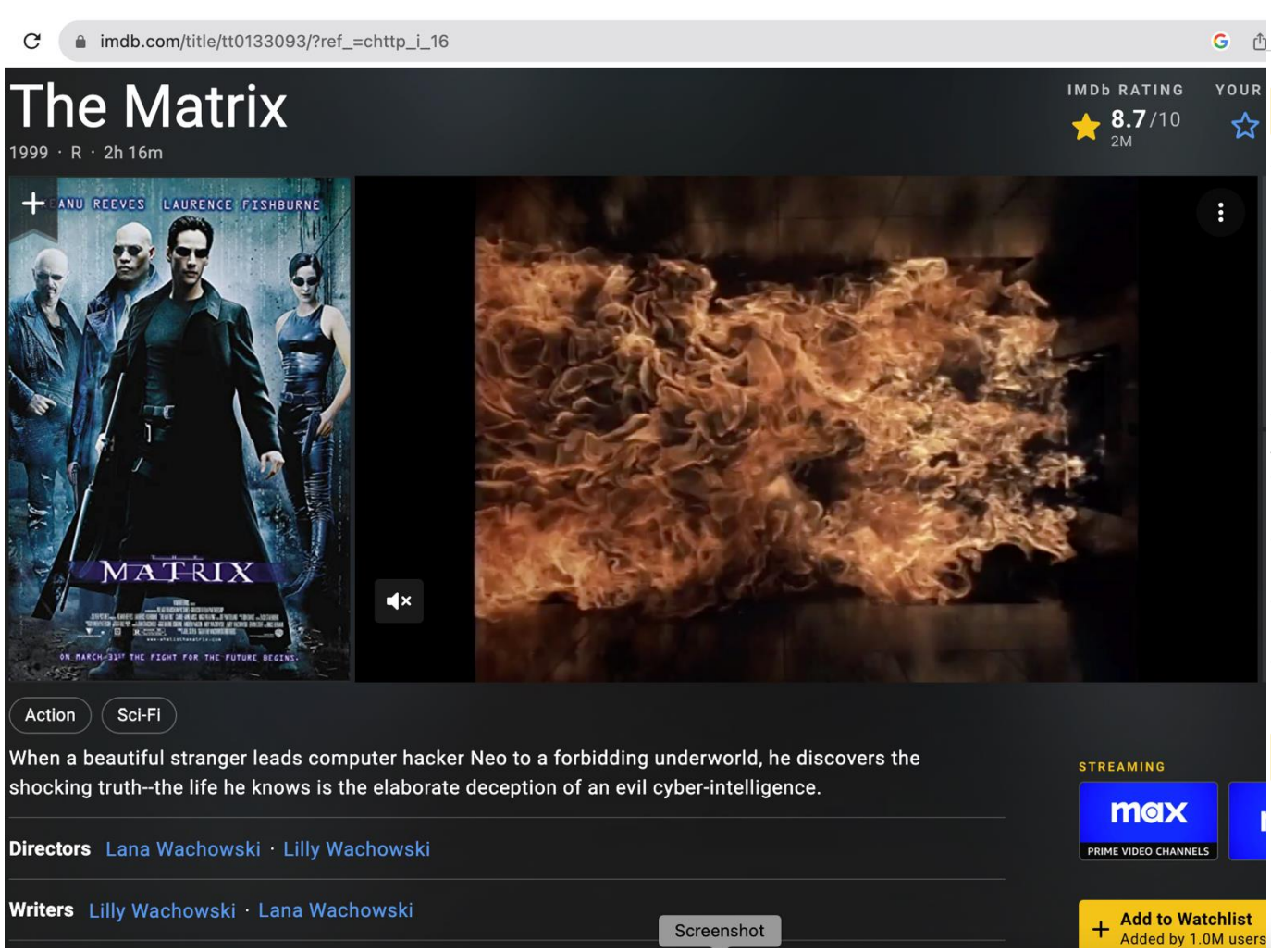
The data science life cycle

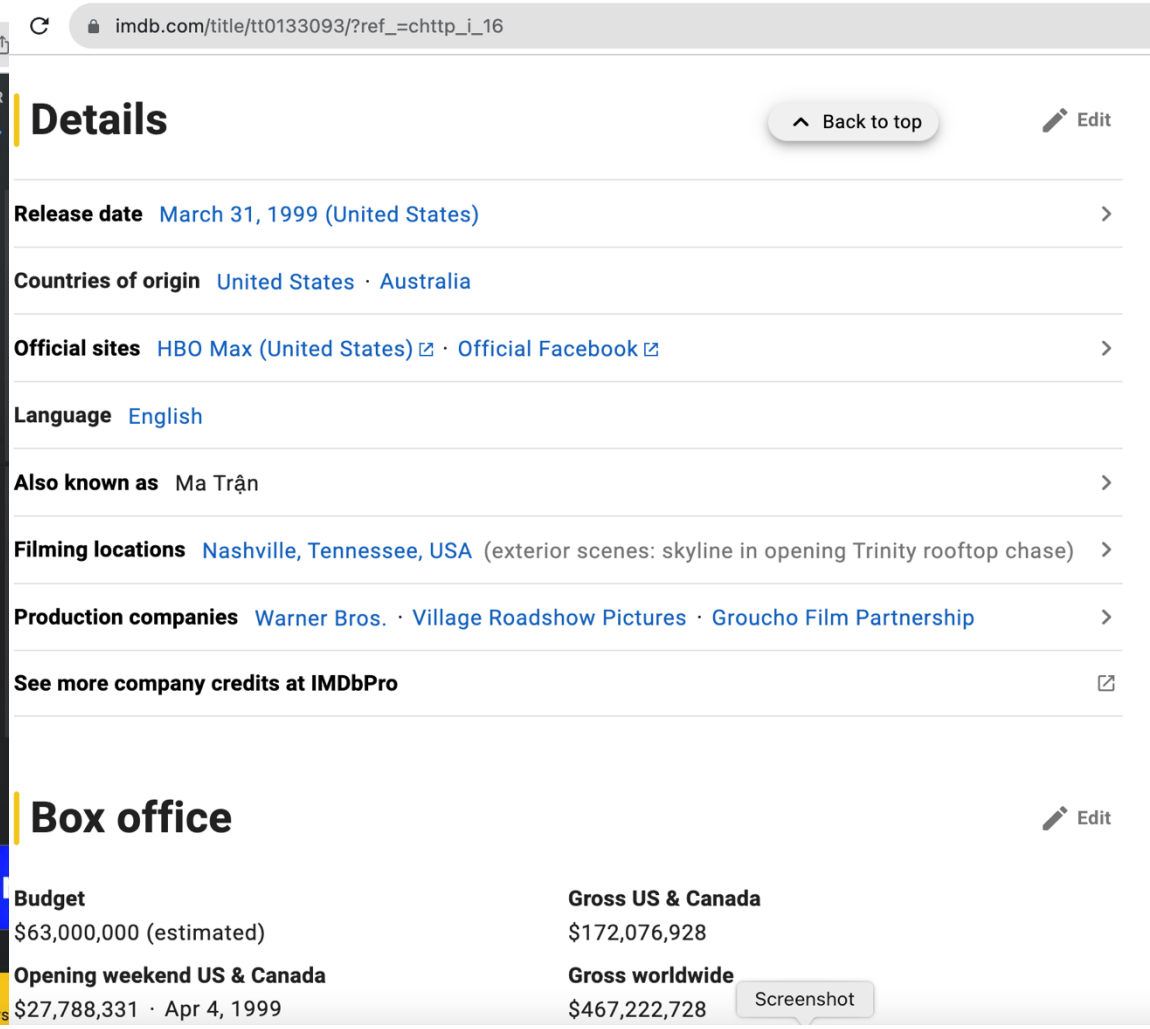
- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



An example

- The IMDb movie database: Movies





An example

- The IMDb movie database: Movies

Jump to

Overview

Born September 2, 1964 · Beirut, Lebanon

Birth name Keanu Charles Reeves

Nicknames The Wall · The One

Height 6' 1" (1.86 m)

Family

Children
No Children

Parents
Samuel Nowlin Reeves
Patric Reeves

Relatives
Kim Reeves (Sibling)
Karina Miller (Half Sibling)
Emma Reeves (Half Sibling)

Trademarks

Intense contemplative gaze

Deep husky voice

Known for playing stoic reserved characters

Friendly, down-to-earth personality

The Matrix (1999)

Edi

Full Cast & Crew

IMDbPro See agents for this cast & crew on IMDbPro

Directed by

Lana Wachowski ... (as The Wachowski Brothers)
Lilly Wachowski ... (as The Wachowski Brothers)

Writing Credits (WGA)

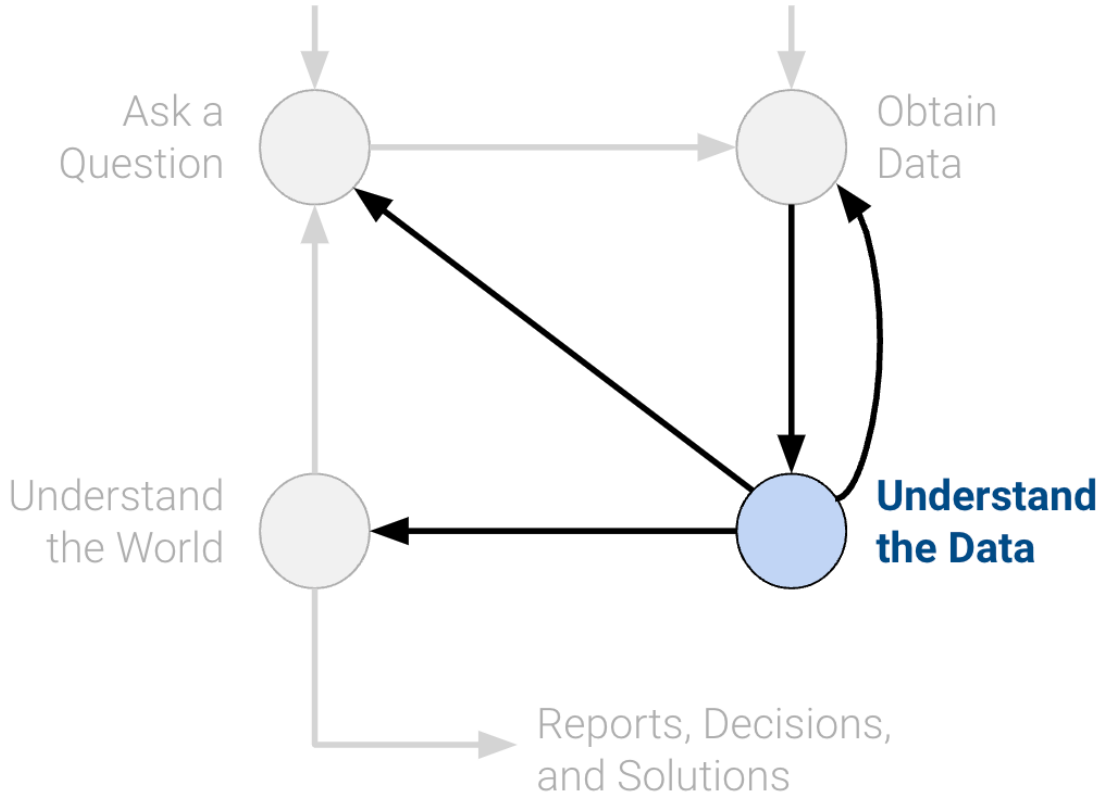
Lilly Wachowski ... (written by) (as The Wachowski Brothers) &
Lana Wachowski ... (written by) (as The Wachowski Brothers)

Cast (in credits order) verified as complete

Keanu Reeves	...	Neo
Laurence Fishburne	...	Morpheus
Carrie-Anne Moss	...	Trinity
Hugo Weaving	...	Agent Smith
Gloria Foster	...	Oracle
Joe Pantoliano	...	Cypher
Marcus Chong	...	Tank
Julian Arahanga	...	Apoc

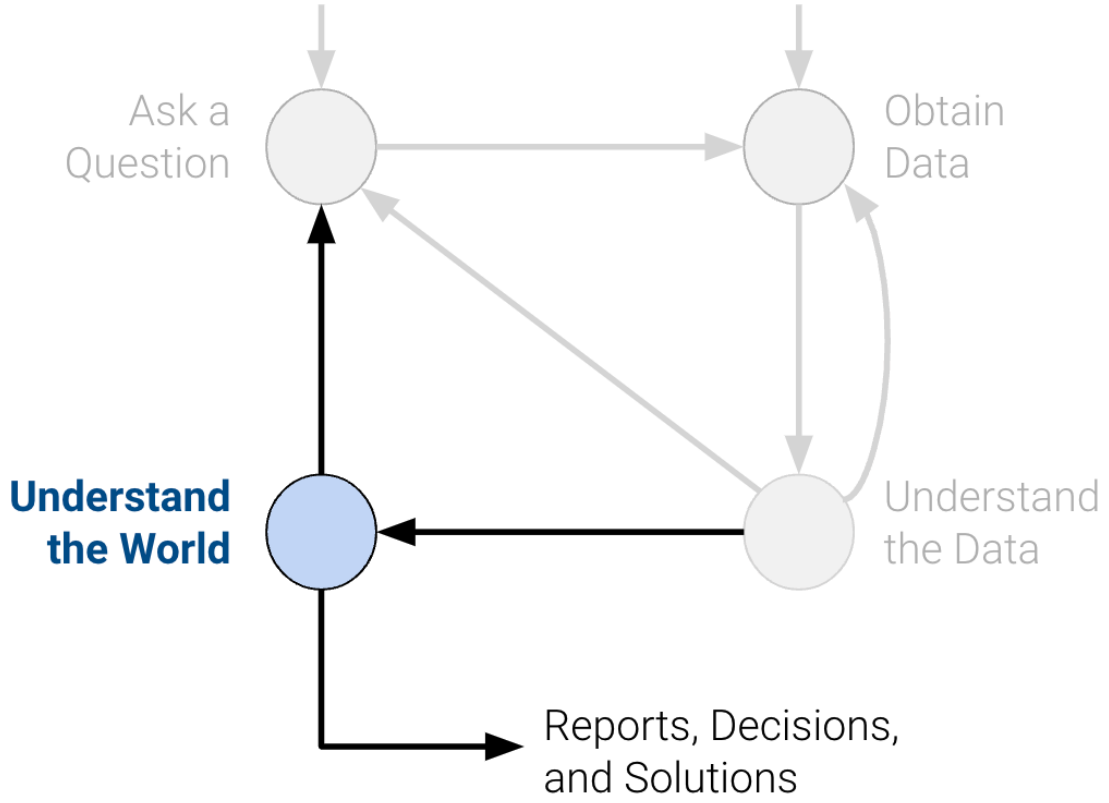
The data science life cycle

- 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?
  - E.g., overrepresented by highly engaged user
- How do we transform the data to enable effective analysis?



The data science life cycle

- 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?



Course Overview

Course Overview: Resources

https://xinchenyu.github.io/csc380/

CSC 380: Principle of Data Science

Overview

This course introduces students to principles of data science that are necessary for computer scientists to make effective decisions in their professional careers. A number of computer science sub-disciplines now rely on data collection and analysis. For example, computer systems are now complicated enough that comparing the execution performance of two different programs becomes a statistical estimation problem rather than a deterministic computation. This course teaches students the basic principles of how to properly collect and process data sources in order to derive appropriate conclusions from them. The course has main components of: basic probability, basic statistics and data wrangling, and basic data analysis using programming libraries.

Logistics info

Time and venue: Tuesday and Thursday 5:00-6:15pm at C E Chavez Bldg 111

- Syllabus
- Gradescope
- D2L course webpage
- Piazza link (access code: `wildcats`)
- Lecture participation self-report form

Specific resources

- gradescope for assignment submission
- Piazza for discussions and Q&A.
- Readings and electronic textbooks
- Lecture slides

Most lecture accompanied by reading

- You are expected to read them

Attendance is required

Course Materials

Textbooks:

- Watkins, J., "An Introduction to the Science of Statistics: From Theory to Implementation"
- Wasserman, L. "All of Statistics: A Concise Course in Statistical Inference." Springer, 2004
- James, G., Witten, D., Hastie, T., & Tibshirani, R. An Introduction to Statistical Learning with Applications in Python. New York: Springer
- Steven S. Skiena, ”The Data Science Design Manual”, Springer, 2017
- Sam Lau, Joey Gonzalez, Deb Nolan, “Learning Data Science”, O’Reilly, 2023

Lecture slides will be shared through course website.

Technologies/Libraries we will use

Expected Skills

- This class will use a fair amount of **math**
  - Probability and Statistics
  - Some basic Calculus and Linear Algebra
  - "Math should be there to aid understanding, not hindering it”-- Hal Daume III
  - Suggestion: attach “physical meanings / pictures” to math equations

- This class will require a fair amount of **coding**
  - Reading in / cleaning / visualizing data
  - Simulating random processes
  - Training and evaluating machine learning models

- Some assignment questions will be **math**, some will be **coding**

Course Overview

Course Objective *Introduction to basic concepts in data science and machine learning.*

Basic Probability	Basic Statistics	Data Handling and Visualization	Machine Learning
random events, variables, distributions, moments	descriptive statistics, estimation, hypothesis testing	reading, cleaning, preprocessing, visualization	predictive models, supervised learning, unsupervised learning

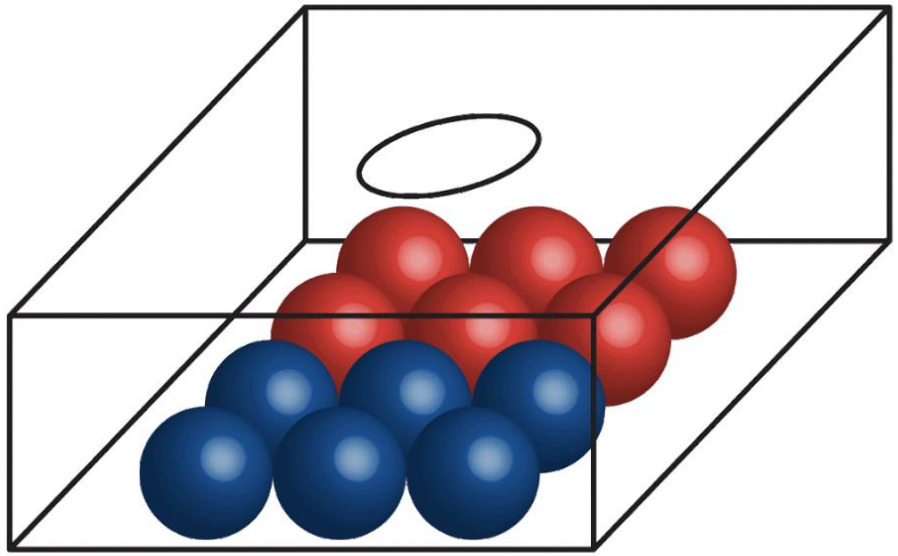
Tentative schedule

> **Note:** All assignments due at 11:59 pm.

Dates	Topics	Assignments	Due Dates
Jan 15	Course Overview		
Jan 20 Jan 22	Basic Data Analysis 1 Probability 1.1	HW1 Out	
Jan 27 Jan 29	Probability 1.2 Probability 2.1	HW2 Out	HW1 Due **JAN 28**
Feb 03 Feb 05	Probability 2.2 Probability 2.3		HW2 Due **FEB 06**
Feb 10 Feb 12	Probability 3.1 Probability 3.2	HW3 Out	
Feb 17 Feb 19	Probability 3.3 Probability 3.4		HW3 Due **FEB 19**
Feb 24 Feb 26	Probability 4.1 Probability 4.2	HW4 Out	
Mar 03 Mar 05	Probability 4.3 Midterm review 1		HW4 Due **MAR 05**
Mar 17 Mar 19	Midterm review 2 **MIDTERM**		
Mar 24 Mar 26	Basic Data Analysis 2.1 Basic Data Analysis 2.2	HW5 Out	
Mar 31 Apr 02	Basic Data Analysis 3.1 Basic Data Analysis 3.2	Project Out	HW5 Due **APR 02**
Apr 07 Apr 09	Basic Data Analysis 3.3 Basic Data Analysis 3.4	HW6 Out	
Apr 14 Apr 16	Basic Data Analysis 4 Basic Statistics 1.1	HW7 Out	HW6 Due **APR 16**
Apr 21 Apr 23	Basic Statistics 1.2 Basic Statistics 2.1		
Apr 28 Apr 30	Basic Statistics 2.2 Basic Data Analysis 5		HW7 Due **APR 30**
May 05	Final review		Project Due **MAY 08**
May 13	**FINAL EXAM**		

- Probability: 12 lectures
- Data analysis: 9 lectures
- Statistics: 4 lectures

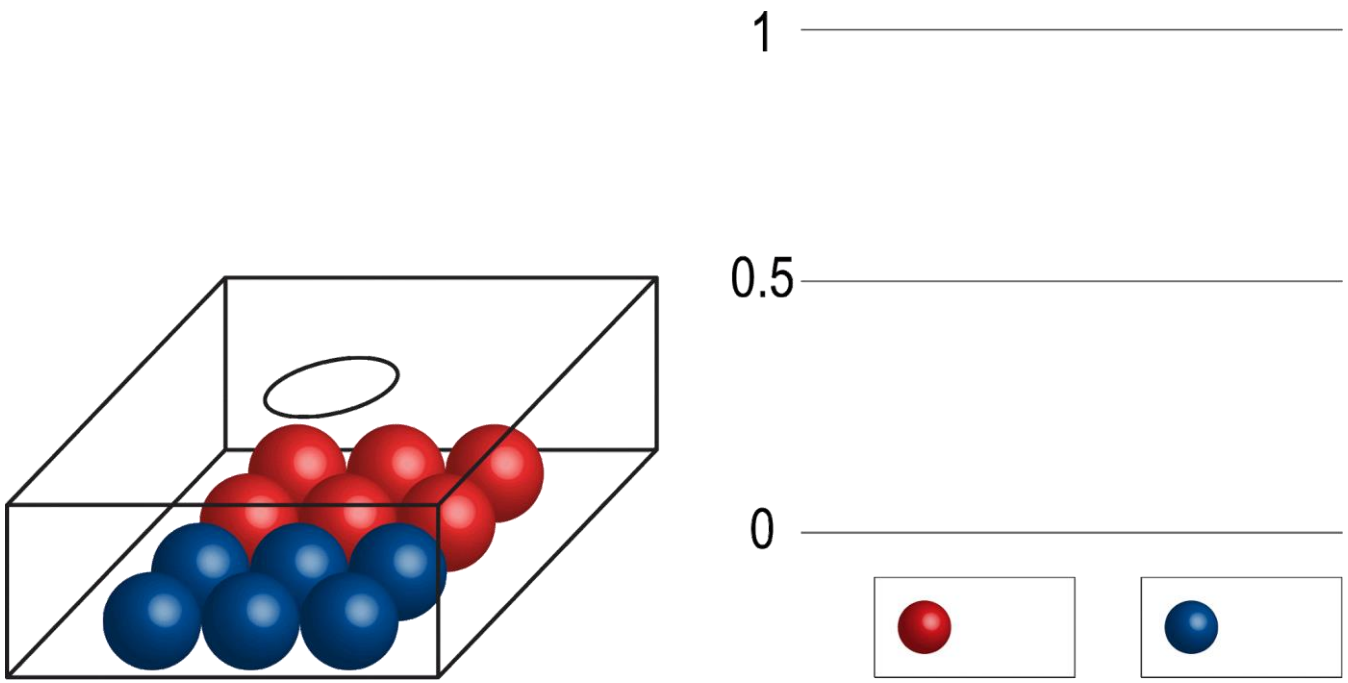
Probability



Probability is about predicting chances using numbers:

- Given that there are 6 red balls and 6 blue balls in the box.
- You pick 1 balls from the box. What is the probability that the ball is red?
- You pick 2 balls from the box. What is the probability that both balls are red?

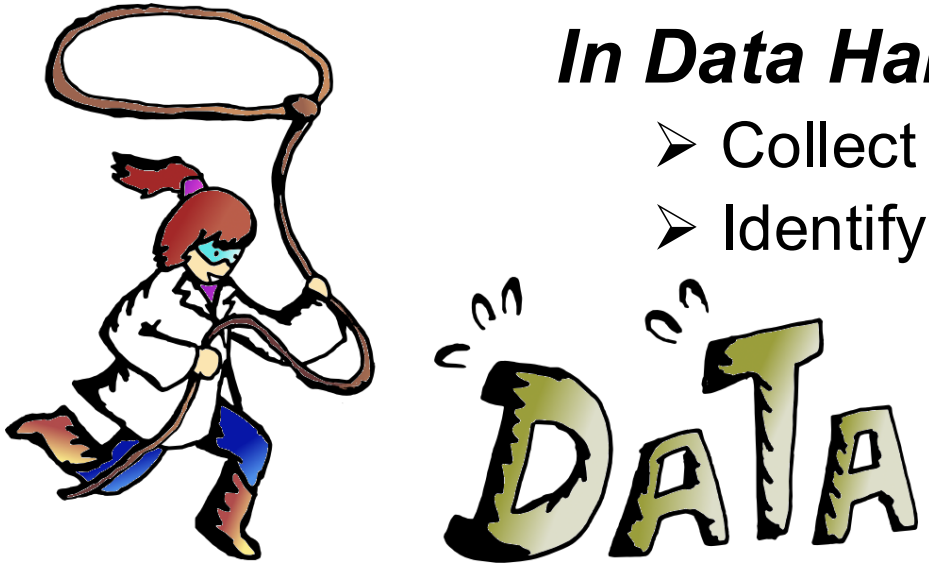
Statistics



Statistics is about turning raw data into insights:

- There are 12 balls in the box, but you don't know how many are red.
- Experiment: pick 1 ball from the box each time then put it back, repeat 100 times.
- Suppose you observe 50 times are red, how many red balls would you guess? Why?

Data Handling and Visualization



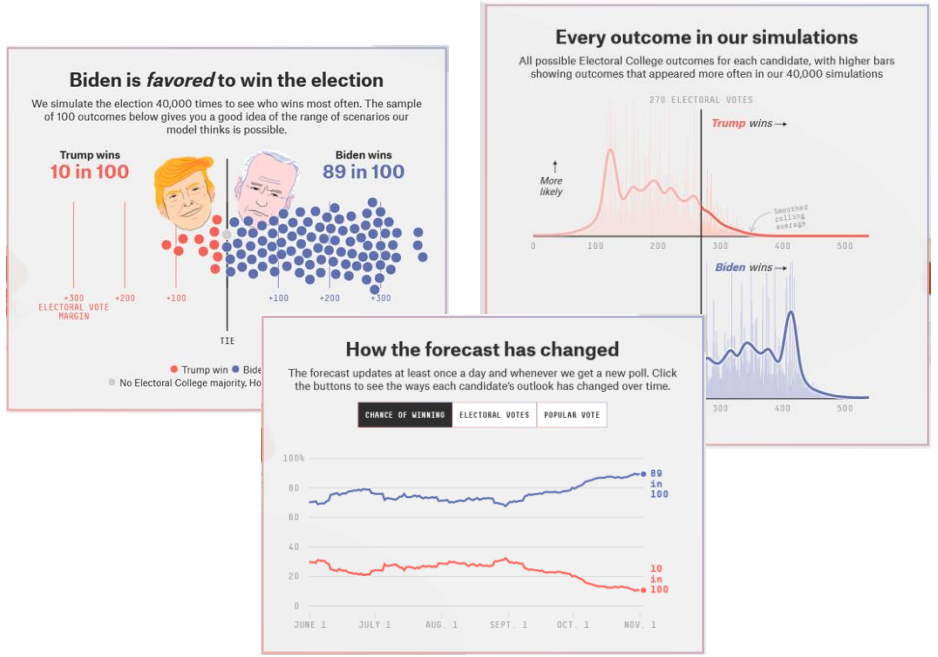
[Image Source: Code A Star]

In Data Handling we will learn to…

- Collect data
- Identify and avoid biased population samples
- Clean data and correct errors
- Transform and preprocess data (***wrangling***)

In Data Visualization we will learn…

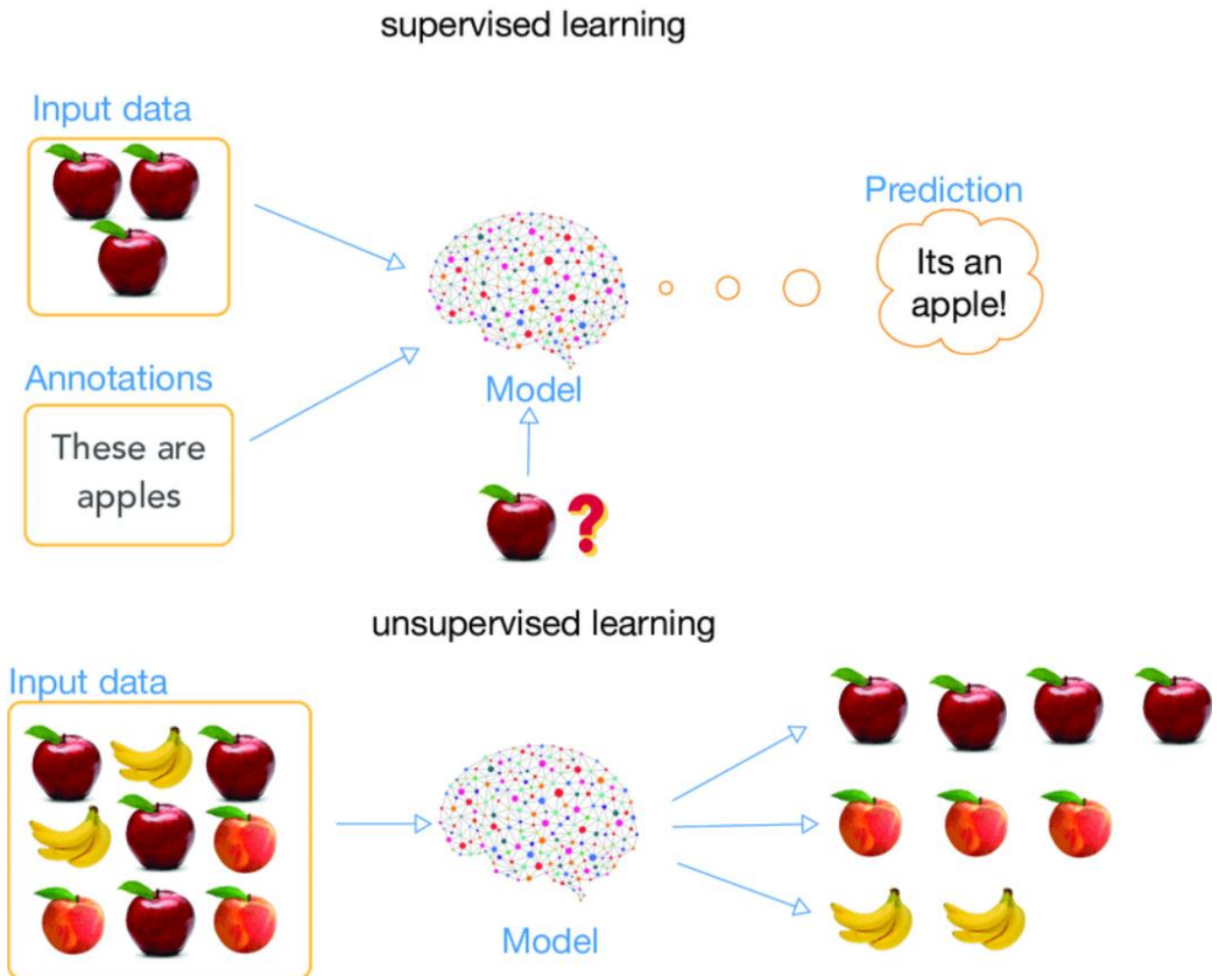
- Why visualization is important
- Exploratory data analysis
- Common forms of visualization
- Pitfalls and gotchas



Machine Learning

In Machine Learning we will learn…

- Principles of prediction
- Unsupervised vs. supervised learning
- Linear and nonlinear models
- Regression and classification



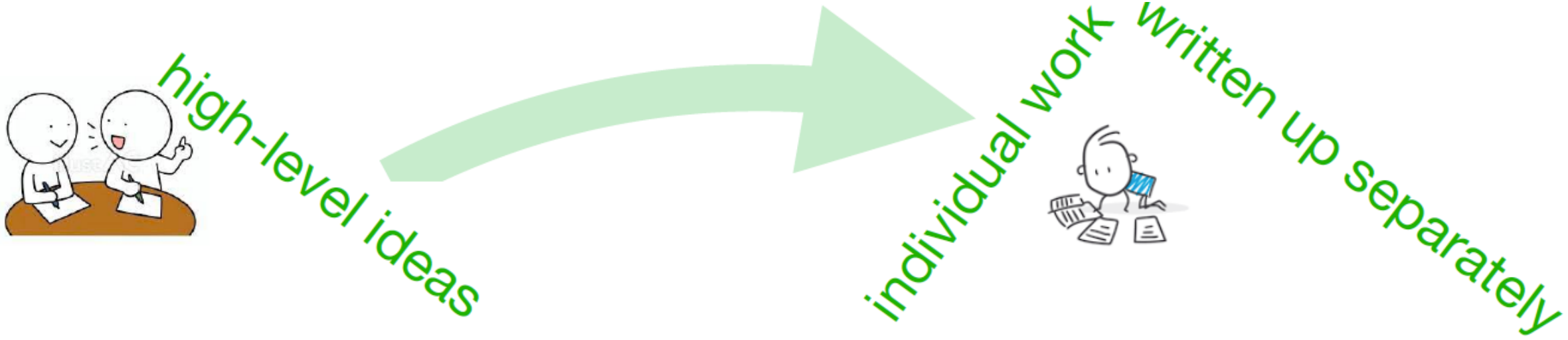
Communication

- Announcements will be made via **D2L**
- Homework submission: **Gradescope** (link: in course website)
- Outside-lecture communications will be through **Piazza** (signup code: in course website)
- The instructor and TAs can also be reached out through email for logistical or personal matters.
- Come to instructor or TAs office hours.

Grading Items

- Assignments (30% total, 5% each)
  - Problem solving + coding, can be done in pairs
  - 7 assignments (lowest one dropped)
- Quizzes (15% total, 1.5% each)
  - 12 quizzes (lowest two dropped), typically on Tuesdays unless otherwise noted (quiz 01 will be next Thursday)
- Project (10%)
  - Groups of size <= 3
- Exams (45% total)
  - 1 midterm (20%) + final (25%)

Homework policy



- If you discuss high-level ideas with a friend mention his/her name in your solution.
- Verbatim copying of solutions from any external source (web search, ChatGPT, etc.) is not acceptable. If you make use of any external source mention the source in your solution and make sure that your solution is your own work.
- Submissions after the due date and time are not accepted.

Bonus points

- 3% bonus: All conditions below must be satisfied.
  - 3 “participation instances” in lectures
  - 3 “participation instances” in piazza
  - 3 appearances in instructor’s or TAs office hours

PAIR UP & DISCUSS

- “participation instance” in lecture: Answer a group (with neighbor) discussion question correctly

- “participation instance” in piazza: A correct answer to a technical question (at most one/month counted)

Important Dates

- Jan 27: last date to self-withdraw without a ‘W’
- Mar 19: midterm
- May 08: project due
- May 13: final exam

Mental Wellbeing

Some occasional stress / depression / anxiety is normal, but sometimes you may need extra help

- Non-emergency UA resources at Counseling & Psych Services Mon-Fri
  - Phone: 520-621-3334
  - Web: https://health.arizona.edu/counseling-psych-services
- Emergency resources in Tucson in this Google Doc

Inclusivity

We want to foster a comfortable and inclusive classroom experience

Please let me know if you feel excluded in any way, e.g.

- Improper use of pronouns
- Microagressions
- Miscellaneous statements / interactions

You can message us on Piazza or discuss in person

Reading Assignments

- Robinson and Nolis, "What is Data Science?" (link from course schedule page)

- 'Probability and statistics cookbook' is a good cheat sheet. Download it from http://statistics.zone/