CSC380: Principles of Data Science

Introduction and Course Overview

Xinchen Yu

Course Instructor

Xinchen Yu

xinchenyu@arizona.edu

Outline

- Data Science Introduction
  - What is data science?

- Course Overview
  - Resources
  - Grading policy
  - What you will learn

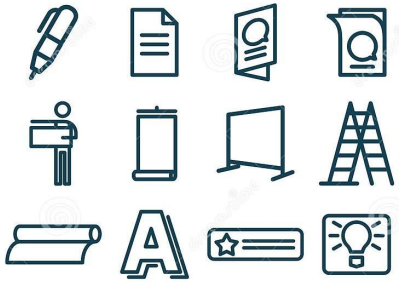
Amazon Sales Dataset

	product_id	product_name	discount_percentage	rating	rating_count	about_product	review_title	review_content
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	64%	4.2	24,269	High Compatibility : Compatible With iPhone 12...	Satisfied,Charging is really fast,Value for mo...	Looks durable Charging is fine tooNo complains...
1	B098NS6PVG	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	43%	4.0	43,994	Compatible with all Type C enabled devices, be...	A Good Braided Cable for Your Type C Device,Go...	I ordered this cable to connect my phone to An...
2	B096MSW6CT	Sounce Fast Phone Charging Cable & Data Sync U...	90%	3.9	7,928	【Fast Charger& Data Sync】-With built-in safet...	Good speed for earlier versions,Good Product,W...	Not quite durable and sturdy,https://m.media-a...
3	B08HDJ86NZ	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	53%	4.2	94,363	The boAt Deuce USB 300 2 in 1 cable is compati...	Good product,Good one,Nice,Really nice product...	Good product,long wire,Charges good,Nice,I bou...
4	B08CF3B7N1	Portronics Konnect L 1.2M Fast Charging 3A 8 P...	61%	4.2	16,905	[CHARGE & SYNC FUNCTION]- This cable comes wit...	As good as original,Decent,Good one for second...	Bought this instead of original apple, does th...

https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset

What is “Data Science”?

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



Examples:

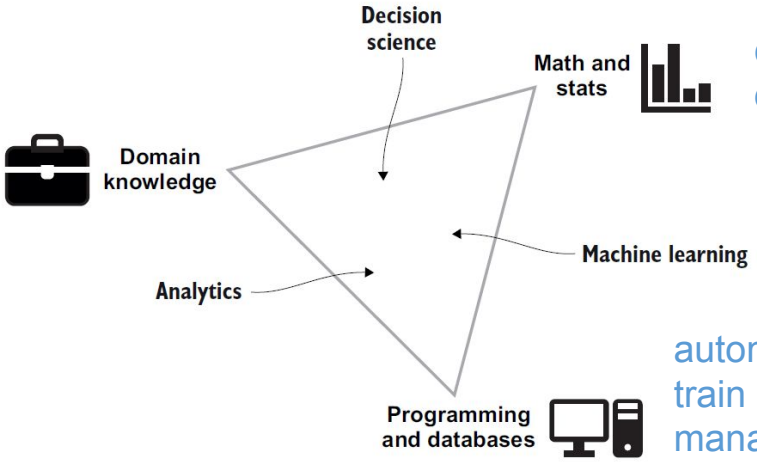
- Do people in college towns tend to buy more notebooks than people in other areas?
- Find out top-10 sales categories for each age group.
- Summarize product reviews w.r.t. product quality, customer service, etc.
- If we recommend pens to users from college town, how much will it increase our revenue?”

What is “Data Science”?

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

amazon:
how customers behave

manufacturing:
how the process works



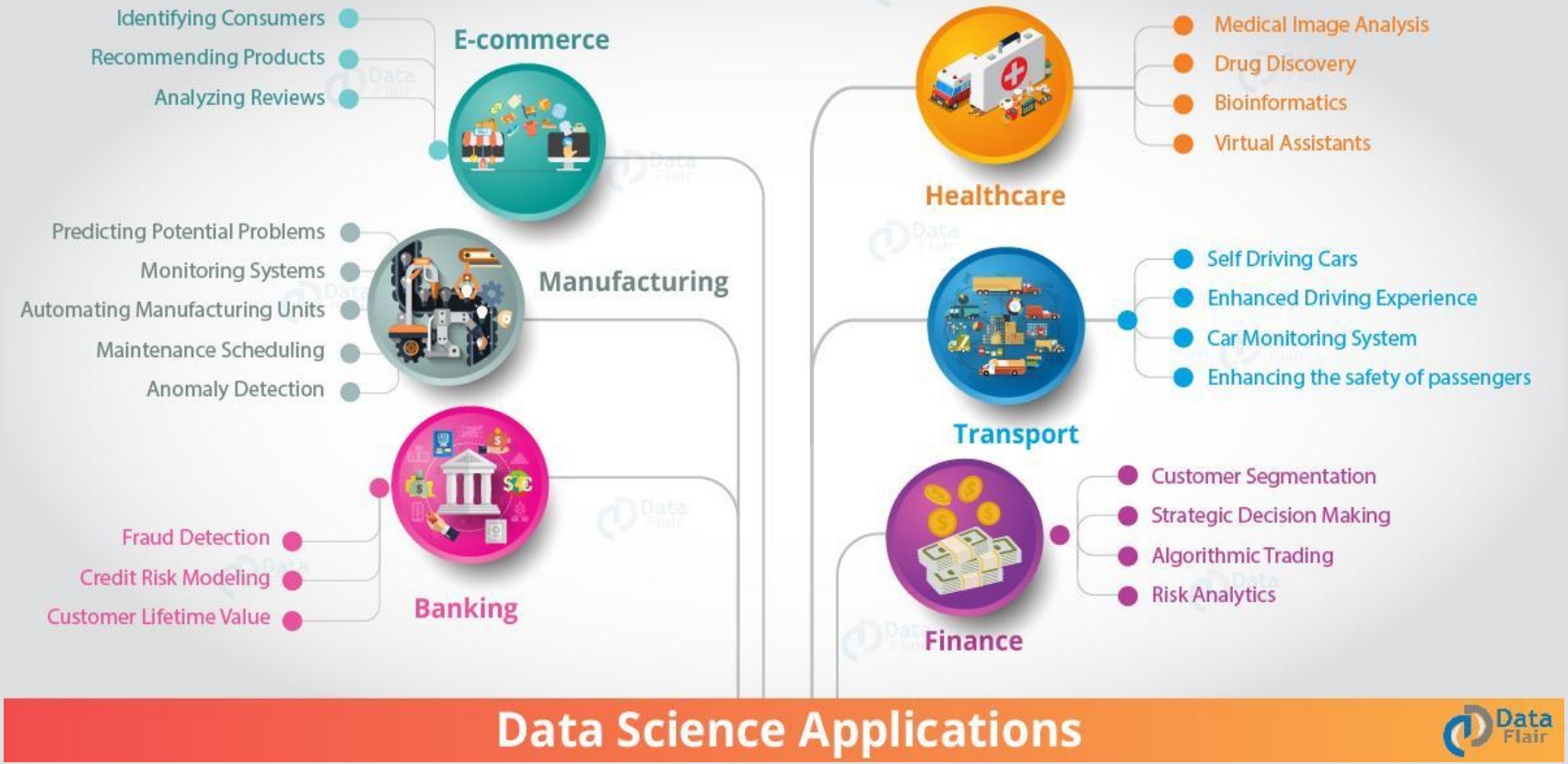
compute average/variance
design statistical models

automate analyses
train models
manage data (store, retrieve)

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

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


Identifying Consumers
Recommending Products
Analyzing Reviews
E-commerce
Predicting Potential Problems
Monitoring Systems
Automating Manufacturing Units
Maintenance Scheduling
Anomaly Detection
Manufacturing
Fraud Detection
Credit Risk Modeling
Customer Lifetime Value
Banking
Medical Image Analysis
Drug Discovery
Bioinformatics
Virtual Assistants
Healthcare
Self Driving Cars
Enhanced Driving Experience
Car Monitoring System
Enhancing the safety of passengers
Transport
Customer Segmentation
Strategic Decision Making
Algorithmic Trading
Risk Analytics
Finance
Data Science Applications
Data Flair

Netflix recommendation systems

Question: How to personalize Netflix as much as possible to a user?

Data sources:

- several billion ratings from its members.
- stream related data (duration, time of playing, day of the week)
- metadata (director, actor, genre, reviews from different platforms)

Algorithms:

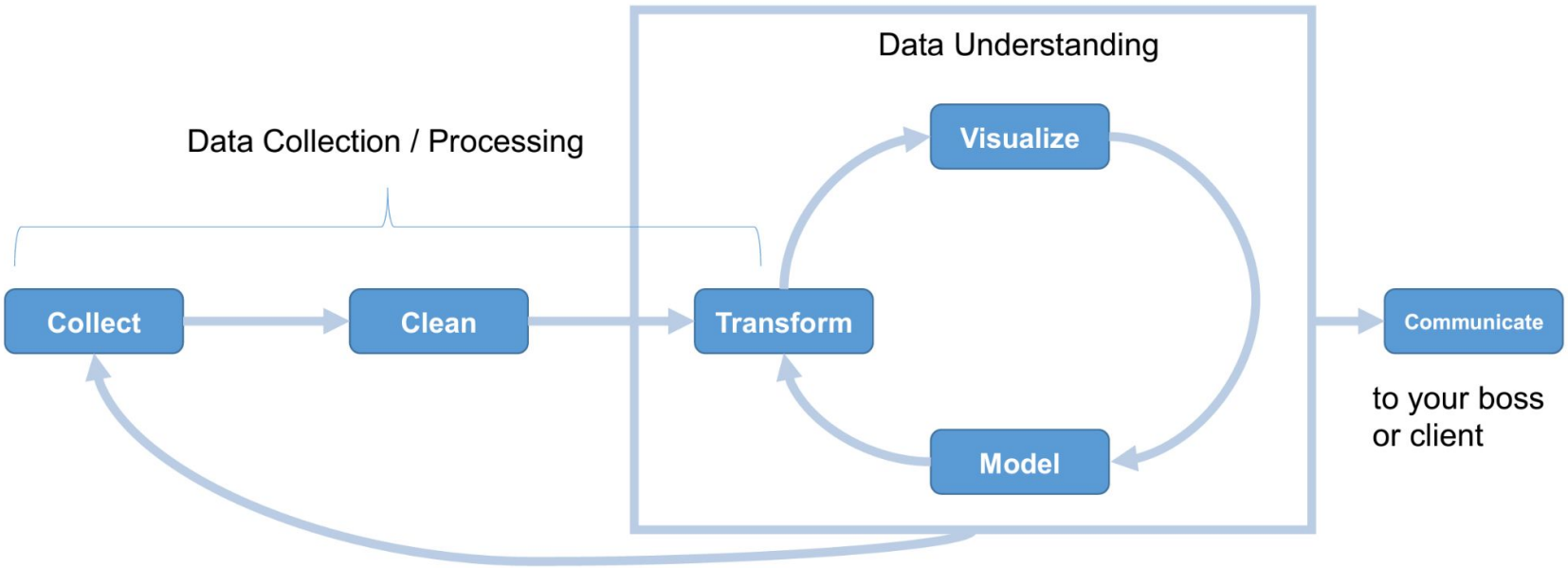
- Top-N Video Ranker
- Continue Watching Ranker
- Video-Video Similarity Ranker

Challenges?

- Parameter tuning
- Whether the effects are due to multiple people sharing the same account / the change in the moods of a person.

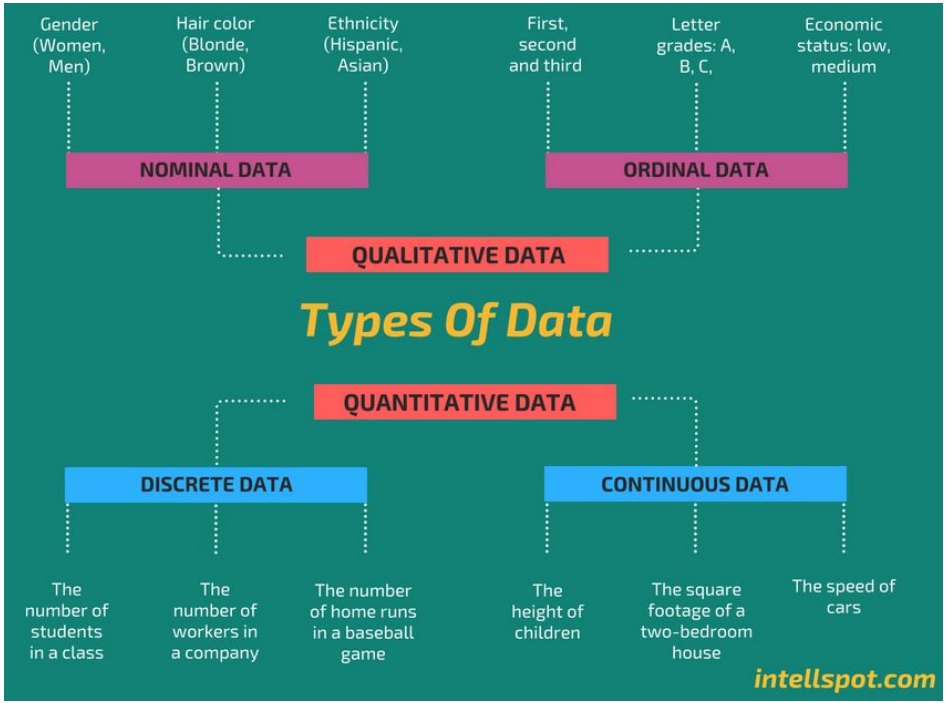
https://towardsdatascience.com/deep-dive-into-netflixs-recommender-system-341806ae3b48

Data Science Workflow



Types of Data

Data come in many forms, each requiring different approaches & models



Natural Language

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Timeseries

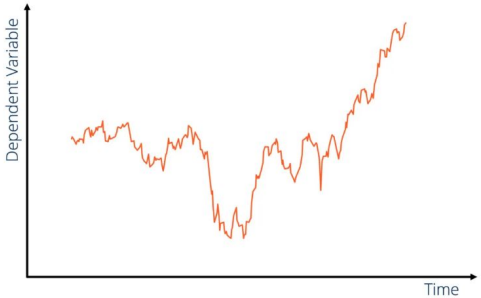
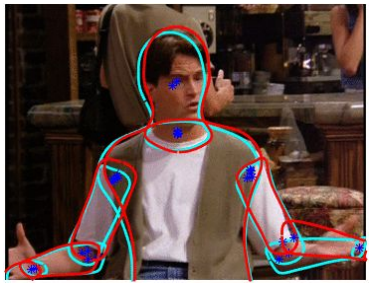


Image / Video



The number of types is endless, these are just some examples

Programming Languages for Data Science

Python and R are both standard for data science these days

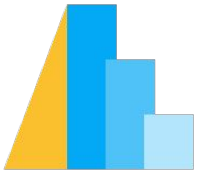
We will use **Python** for this course since you should already know it

julia

↑ **getting popular nowadays**

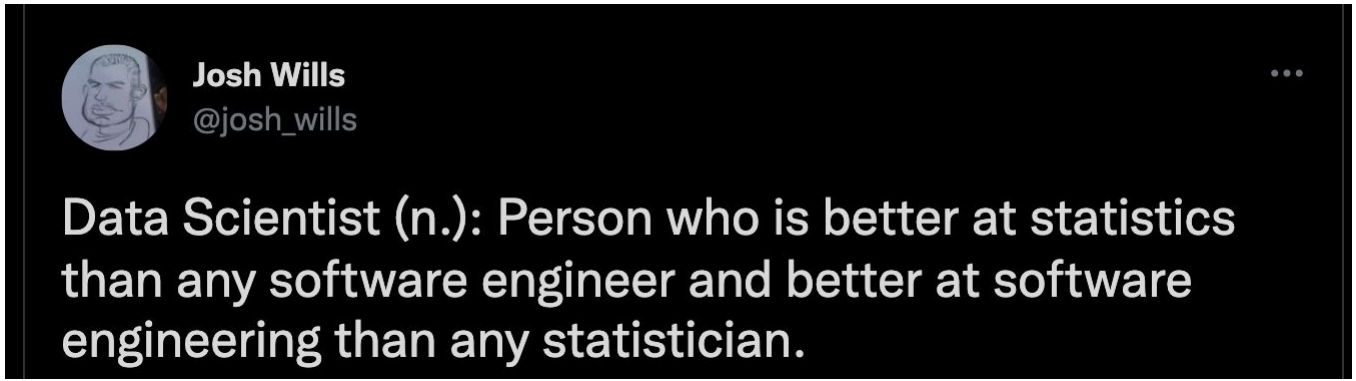
Python Packages Covered

Other Useful Python Packages



Altair

Who is a Data Scientist?



So, you should hone your statistical skills and your value will increase in the job market!!

Data Science Job Market

A search of "data scientist" jobs in the US shows…

Many job options available

- Indeed: 42,000+ jobs
- Glassdoor: 24,000+ jobs
- LinkedIn: 63,000+ jobs

2023's #2 best job in America, according to Indeed.com
(after Full Stack Engineer)

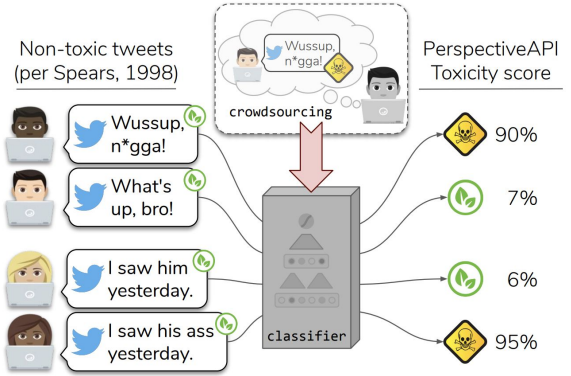
Lucrative pay (Glassdoor)

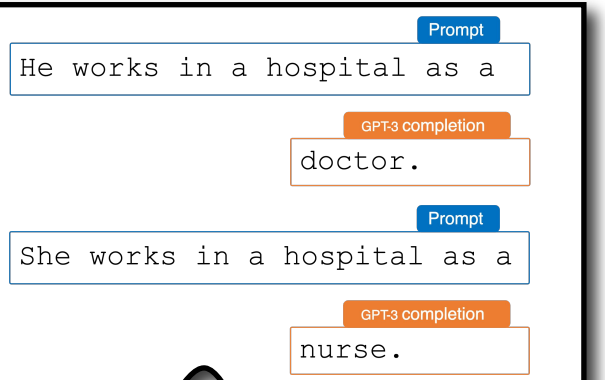
Bad Data Science & Statistics

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

By James Vincent | Jan 12, 2018, 10:35am EST

THE VERGE





Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

REUTERS

Course Overview

Course Overview: Resources

https://xinchenyu.github.io/csc380-spring24/

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 three main components: data analysis, machine learning, and a project where students apply the concepts discussed in class to a substantial open-ended problem.

Logistics info

Time and venue: Tuesday and Thursday 3:30-4:45pm

- Syllabus
- Piazza link Access code: wildcats
- Gradescope Entry code: 8EZRBY (NB: Please make sure your gradescope email address is the same as the one you have on D2L.)
- D2L course webpage: lecture video recordings will be at "UA Tools" -> "Zoom" (NB: Zoom links are for **recordings only** and are not for live-streaming lectures.)

Specific resources

- gradescope for assignment submission
- Piazza for discussions and Q&A.
- Readings and electronic textbooks
- Lecture slides (posted before class)

Every lecture accompanied by reading

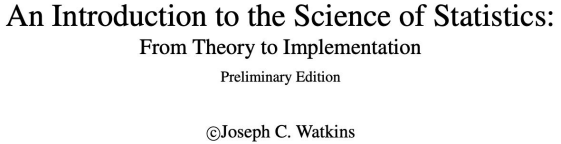
- These will not be graded but are recommended

Attendance is required

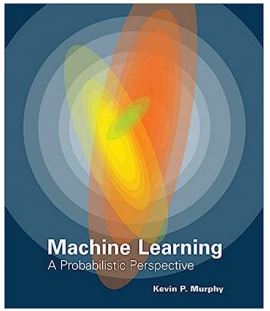
Recordings will be available **after the class**.

Textbooks

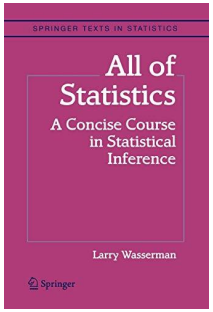
WJ: Watkins, J., "An Introduction to the Science of Statistics: From Theory to Implementation" (https://www.math.arizona.edu/~jwatkins/statbook.pdf)



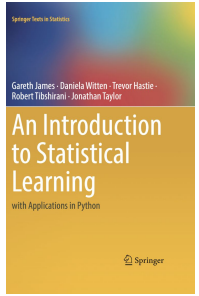
MK: Murphy, K. "Machine Learning: A Probabilistic Perspective." MIT press, 2012 (UA Library)



WL: Wasserman, L. "All of Statistics: A Concise Course in Statistical Inference." Springer, 2004 (UA Library)



ISL: James, G., Witten, D., Hastie, T., & Tibshirani, R. An introduction to statistical learning with Applications in Python. New York: Springer (https://www.statlearning.com/)



Expected Skills

- This class will use a fair amount of **math**
  - Probability and Statistics
  - Some basic Calculus and Linear Algebra
  - These are not required background for the course, but you will learn key concepts in the class.

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

- Some assignments will be **math**, some will be **coding**

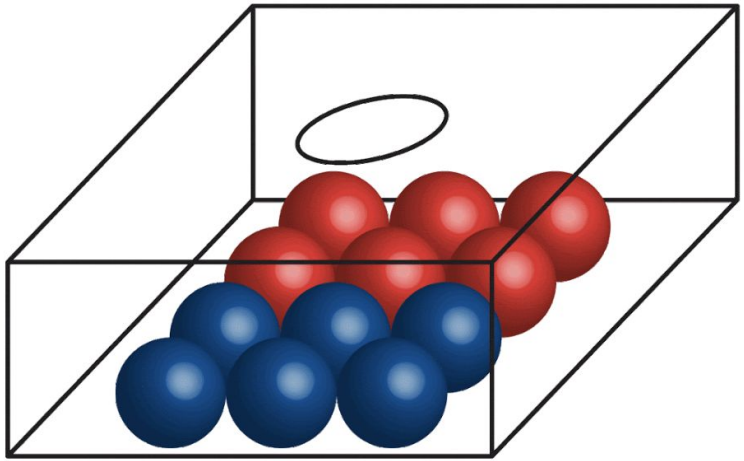
Course Overview

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

Probability and Statistics	Data Handling and Visualization	Machine Learning
Random events / variables, distributions / densities, moments, descriptive stats, estimation	Reading & cleaning, transformation & preprocessing, visualization	Predictive models, supervised learning, unsupervised learning, model checking

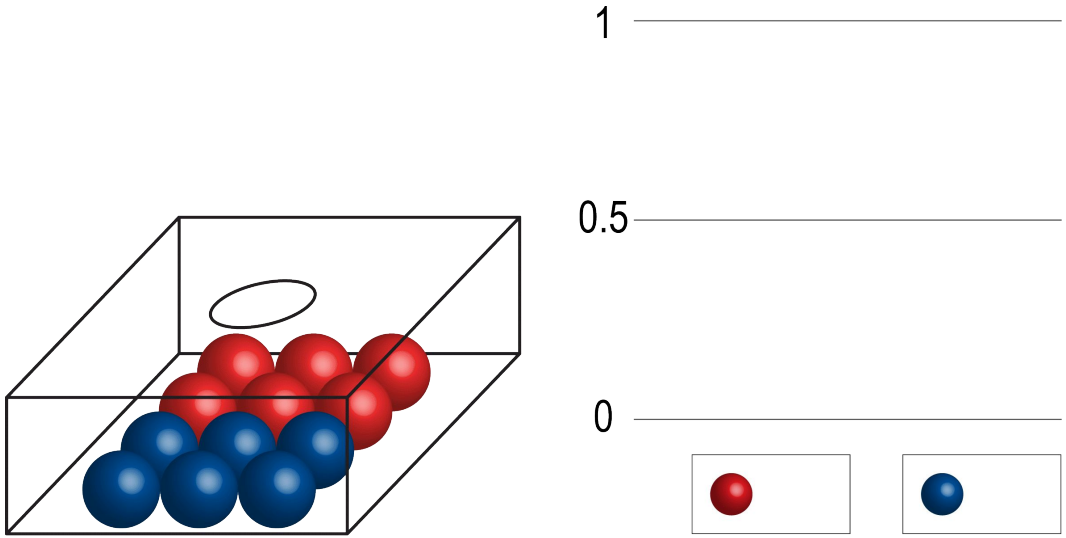
↑ more on this in CSC 480/580

Probability



- You already know there are 6 red balls and 6 blue balls in the box.
- You pick 3 balls from the box.
- What is the probability that all of the 3 balls are red?

Statistics

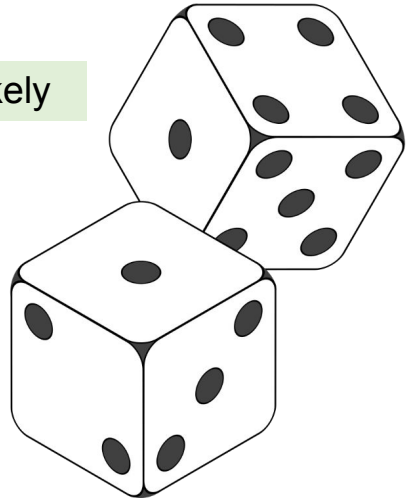


- You know there are 12 balls, but you don’t know how many are red balls.
- Intuition: pick 1 ball from the box each time then put it back, repeat 100 times.
- Observe 50 times are red, so you believe the box has 6 red balls.
- Why?

Probability and Statistics

Suppose we roll <u>two fair dice</u>…

fair die: each side is equally likely



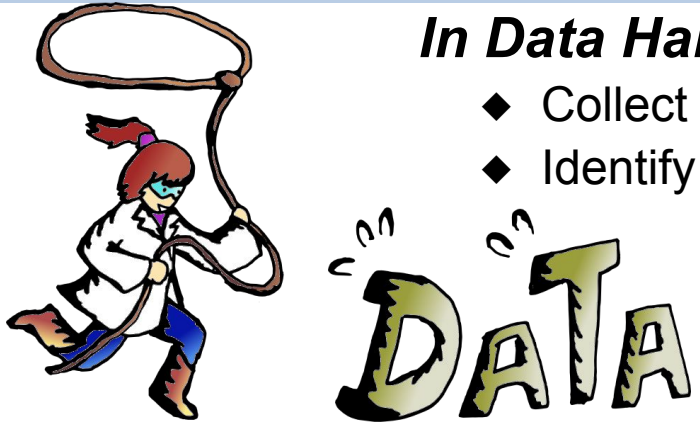
- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?

*… this is a **random trial** or **random process**.*

We will learn how to…

- Mathematically formulate outcomes and their probabilities?
- Describe characteristics of random processes
- Estimate unknown quantities (e.g. are the dice actually fair?)
- Characterize the uncertainty in random outcomes
- Identify and measure dependence among random quantities

Data Handling and Visualization



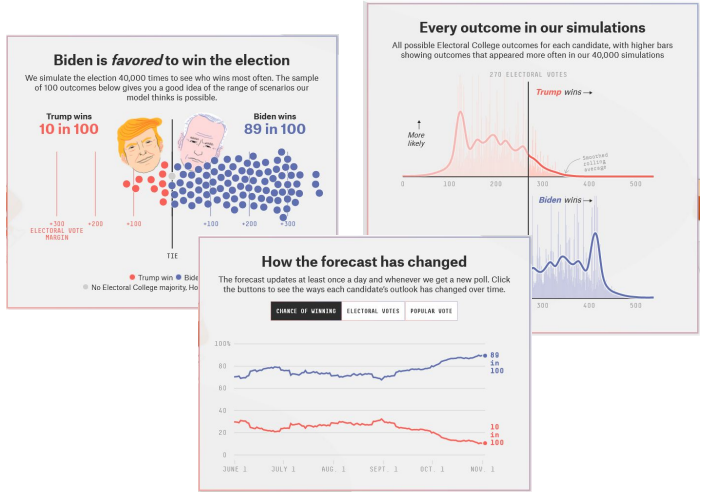
[Image Source: Code A Star]

In Data Handling we will learn to…

- Collect data
- Identify and avoid biased population samples
- Clean data and correct errors
- Transform and preprocess data

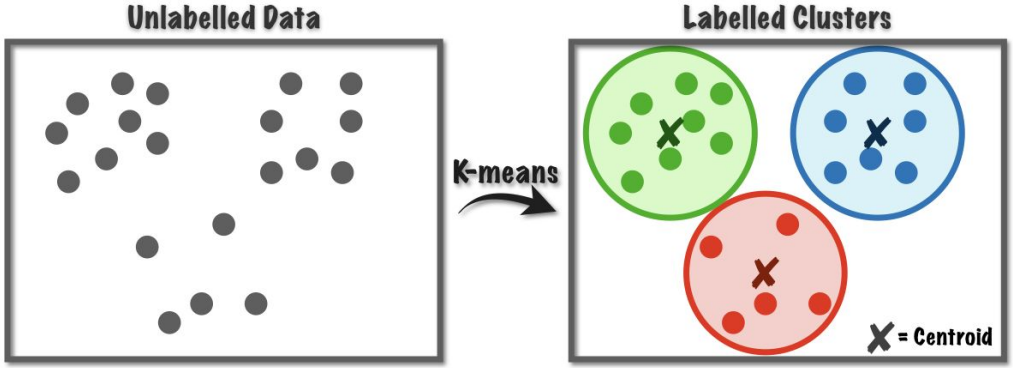
In Data Visualization we will learn…

- Why visualization is important
- Exploratory data analysis
- Common forms of visualization



Machine Learning

How to use data to learn underlying patterns and predict unknowns?



In Machine Learning we will learn…

- Principles of prediction
- Proper partitioning of training / validation / test data
- Unsupervised vs. supervised learning
- Linear and nonlinear models

[Image Source: Towards Data

Assignments / Exams / Grading

8 Homeworks + Midterm + Project + Final Exam

Homeworks

- Homeworks will be due in 8 days: e.g., out on Thursday, due on next Friday.
- You can do HW1, 2 and 4 in pairs, but you must contribute equally for each question if working in pairs
- Grading will be available in 7 days excluding weekends/holidays.
- The one HW with the lowest score will be dropped

Grading Breakdown

- Assignments: 35% (5% each)
- Midterm: 25%
- Project: 10%
- Final Exam: 25%
- Participation: 5%

First assignment out next Thursday

Late Policy

Late submissions impact other students, delay grading, and delay solutions

No late submission policy

- Late submissions are not accepted, period.
- Strongly recommend that you plan to submit your work a day earlier.

Project

- It is a previous competition.
- Data will be given.
- A flexible project. You will answer some open questions.
- You will get a change to try out various ML algorithms and get high accuracy.

Communication

- Announcements will be made via D2L
- Homework submission: **gradescope** (see course website for the link)
- **Piazza** (see course website for the link): we highly encourage that you ask and answer questions among yourselves.
  - We will chime in often.
  - You can also ask questions in piazza directly to us if it is personal.
  - Otherwise, please make the question as a public post so other students can benefit from it.
- Do NOT email me or TAs
  - If you have a private question, make it a private Piazza post

Office Hours

- Office hours will be held both in person and zoom, and the zoom link will be accessible via D2L
- 1hr by the instructor and each TA, once a week.
- The office hour schedule will be announced by next week.
- If you have a conflict with the schedule, let me know (Piazza)

Full Course Schedule (Tentative)

Dates	Topics	Homework	Notes
Jan 11	Course Overview		
Jan 16 Jan 18	Probability 1 Probability 2	HW1 Out	
Jan 23 Jan 25	Probability 3 Probability 4		HW1 Due JAN 26
Jan 30 Feb 1	Probability 5 Probability 6	HW2 Out	
Feb 6 Feb 8	Data processing and visualization 1 Data processing and visualization 2	HW3 Out	HW2 Due FEB 7
Feb 13 Feb 15	Data processing and visualization 3 Statistics 1		HW3 Due FEB 16
Feb 20 Feb 22	Statistics 2 Statistics 3	HW4 Out	
Feb 27 Feb 29	Statistics 4 Midterm review	HW5 Out	HW4 Due FEB 28
Mar 12 Mar 14	Statistics 5 MIDTERM		HW5 Due MAR 15

No in-person lectures on Feb 8 and 13.
Lectures recording will be uploaded to D2L.

Mar 19 Mar 21	Basics of predictive modeling and classification 1 Basics of predictive modeling and classification 2	HW6 Out	
Mar 26 Mar 28	Basics of predictive modeling and classification 3 Linear models 1		HW6 Due MAR 29
Apr 2 Apr 4	Linear models 2 Linear models 3	HW7 Out	
Apr 9 Apr 11	Linear models 4 Nonlinear models 1	HW8 Out	HW7 Due APR 10
Apr 16 Apr 18	Nonlinear models 2 Nonlinear models 3	Project Out	HW8 Due APR 19
Apr 23 Apr 25	Clustering Course wrap-up 1		
Apr 30	Course wrap-up 2		Project Due MAY 2
May 8	FINAL EXAM		

Important Dates

1/23/2024	Last day to drop without a grade of W (withdraw)
3/14/2024	Midterm
4/16/2024	Final project out
5/2/2024	Final project due
5/8/2024	Final Exam

Academic Integrity

Assignments are to be done independently,
unless explicitly marked as a collaborative homework.

If I or the TA suspects you of having cheated

- You will be notified immediately
- We will have a conference where you can plead your case
- If we are not swayed then **you will get an F grade**, period.

To avoid any unconscious cheating, you must write down **who you have worked with** and **to what degree** you got help, outside your group.

Bottom line: don't cheat

Reading Assignments

- Robinson and Nolis, "What is Data Science?" (link from course week 1 page)

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