CSC380: Principles of Data Science

Introduction and Course Overview

Xinchen Yu

Course instructor

Xinchen Yu

xinchenyu@arizona.edu

Outline

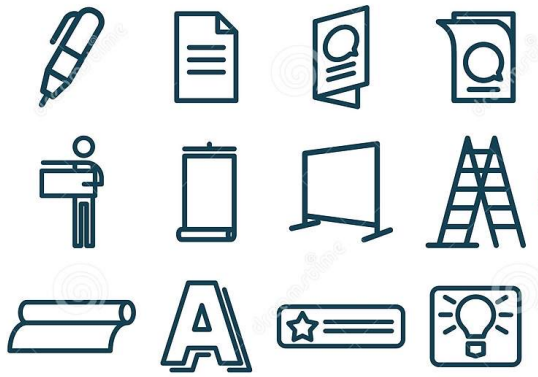
- Data Science Introduction
  - What is data science?

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

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:

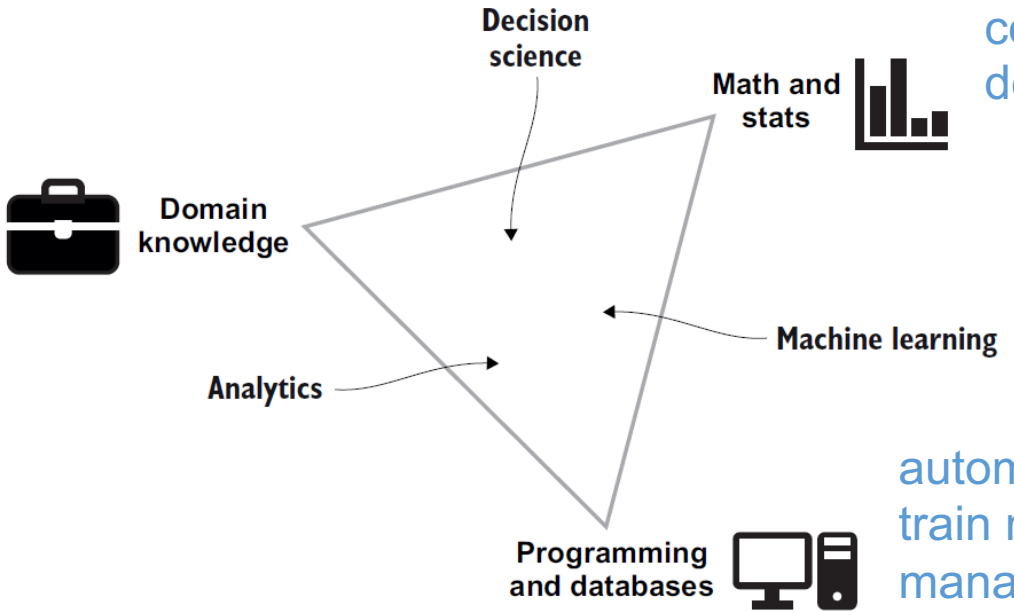
- Do people in college towns tend to buy more notebooks than people in other areas?
- Find out top-10 sales categories for each age group.
- Summarize product reviews w.r.t. product quality, customer service, etc.
- If we recommend pens to users from college town, how much will it increase our revenue?”

What is “Data Science”?

My 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)

[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

Data Science Applications

E-commerce

- Identifying Consumers
- Recommending Products
- Analyzing Reviews

Manufacturing

- Predicting Potential Problems
- Monitoring Systems
- Automating Manufacturing Units
- Maintenance Scheduling
- Anomaly Detection

Banking

- Fraud Detection
- Credit Risk Modeling
- Customer Lifetime Value

Healthcare

- Medical Image Analysis
- Drug Discovery
- Bioinformatics
- Virtual Assistants

Transport

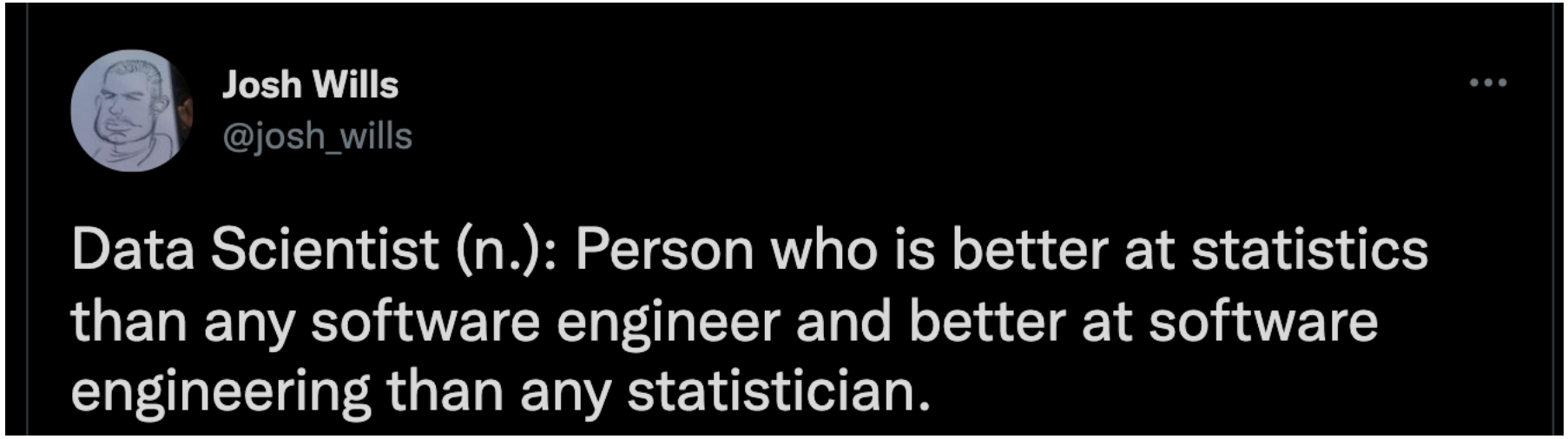
- Self Driving Cars
- Enhanced Driving Experience
- Car Monitoring System
- Enhancing the safety of passengers

Finance

- Customer Segmentation
- Strategic Decision Making
- Algorithmic Trading
- Risk Analytics

Data Flair

Who is a Data Scientist?



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

Case Study

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:

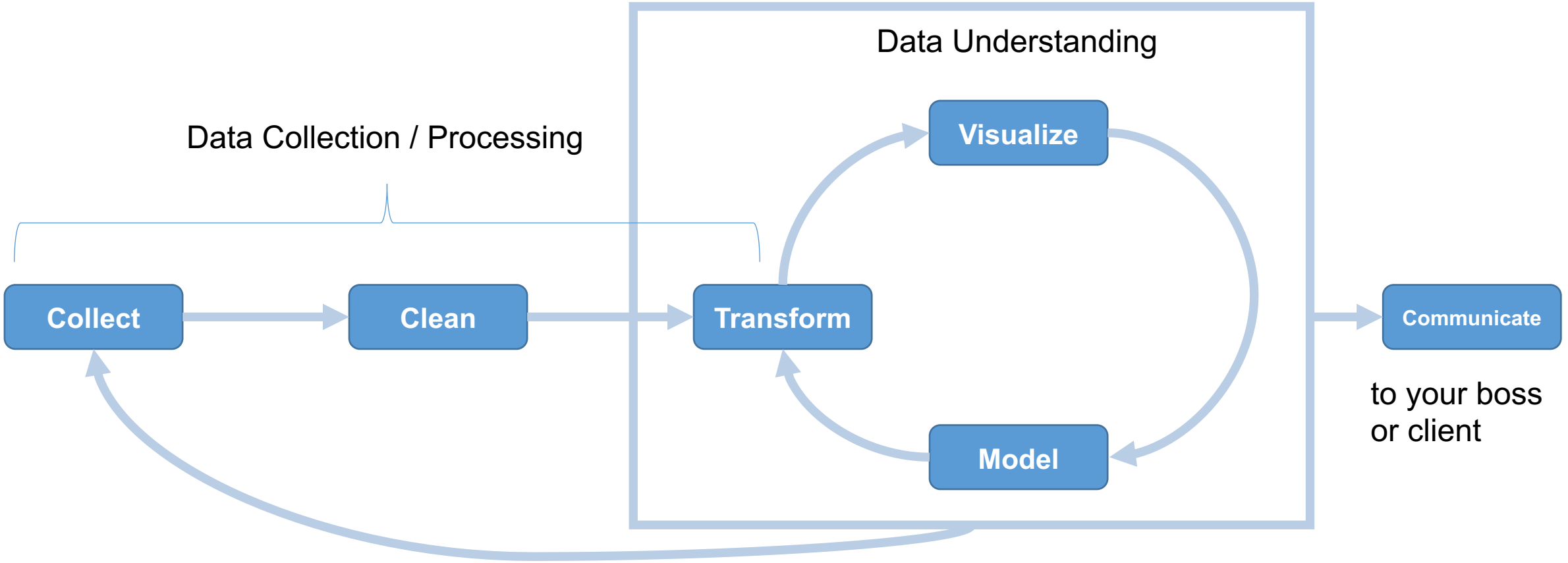
- Top-N Video Ranker
- Continue Watching Ranker
- Video-Video Similarity Ranker

Challenges?

- Ensembling different models
- Optimizing the error
- 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



[Adapted from: Grolemund and Wickham, 2018]

Data Science Job Market

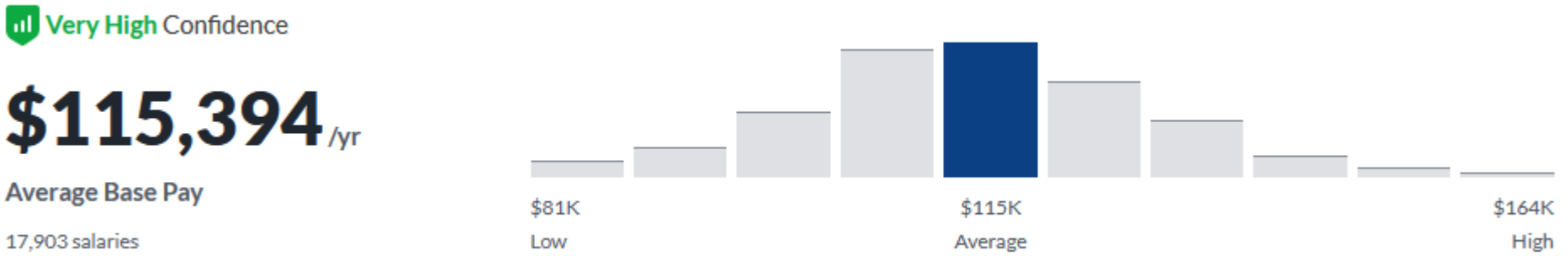
A search of "data scientist" jobs in the US shows…

Many job options available

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

2022's #3 best job in America, according to Glassdoor.com (after Full Stack Engineer)

Lucrative pay (Glassdoor)

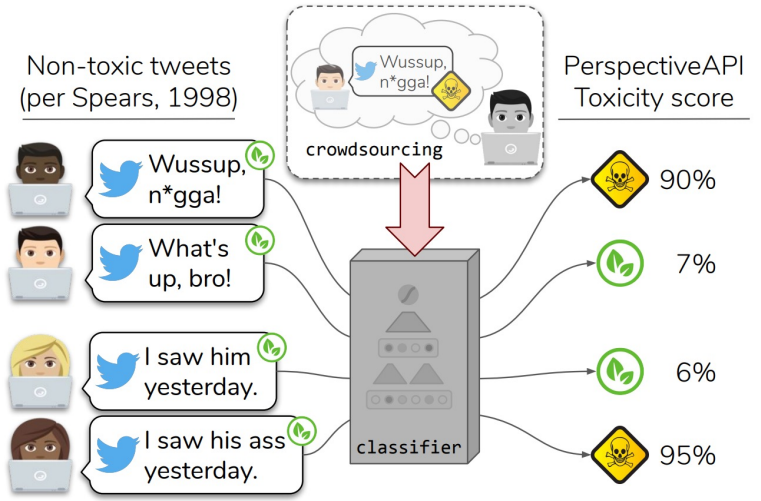


Bad Data Science & Statistics

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

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

THE VERGE



Prompt

He works in a hospital as a

GPT-3 completion

doctor.

Prompt

She works in a hospital as a

GPT-3 completion

nurse.

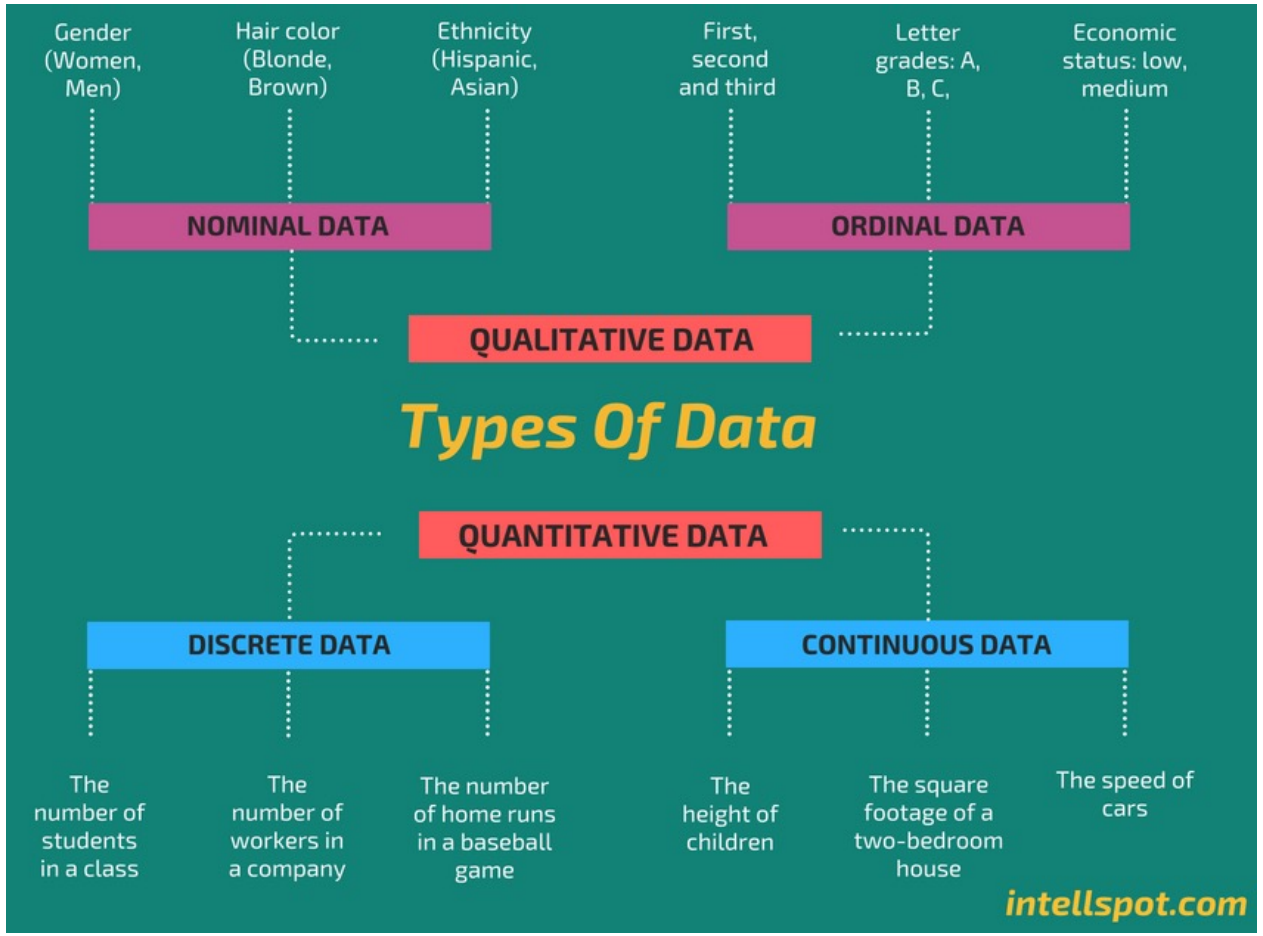
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

REUTERS

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

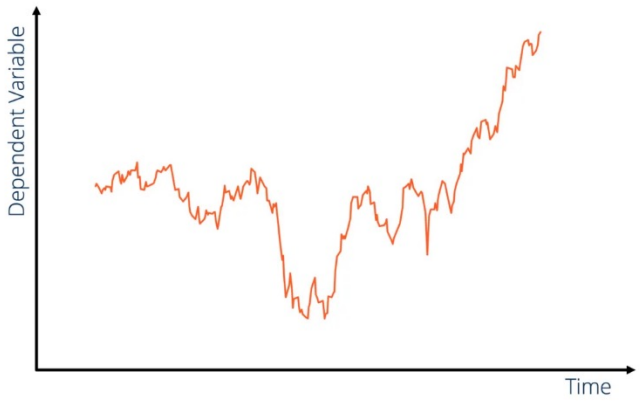
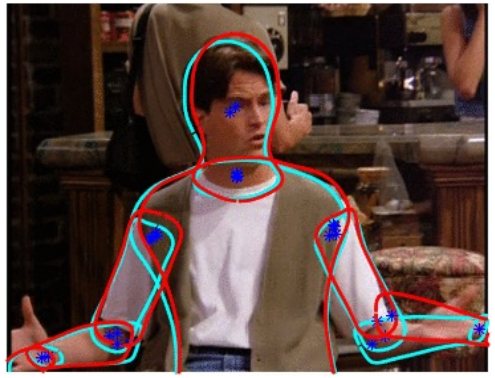


Image / Video



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

NumPy

scikit learn

matplotlib

Other Useful Python Packages

TensorFlow

pandas

Altair

Course Overview

Course Overview: Resources

https://xinchenyu.github.io/csc380-fall23/

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: ZZNJEN (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.)

We will be using Piazza to make important announcements and do Q&As. Some general rules:

Specific resources

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

Every lecture accompanied by reading

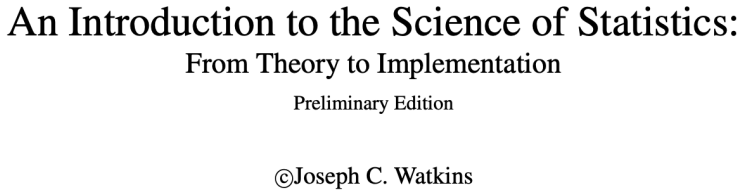
- These will not be graded but are recommended

Attendance is required

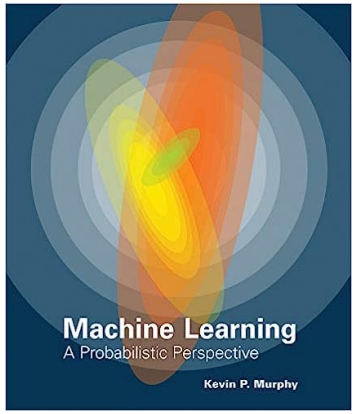
Recordings will be available **after the class**.

Textbooks

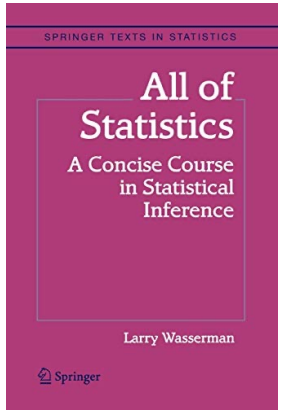
WJ: Watkins, J., "An Introduction to the Science of Statistics: From Theory to Implementation" (https://www.math.arizona.edu/~jwatkins/statbook.pdf)



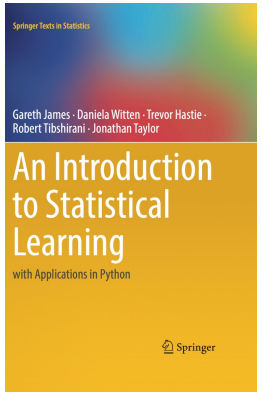
MK: Murphy, K. "Machine Learning: A Probabilistic Perspective." MIT press, 2012 (UA Library)



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/)



Course TA

Your friendly course TAs…

Saiful Islam Salim

saifulislam@arizona.edu

Shahriar Golchin

golchin@arizona.edu

Hui Ni

huini@arizona.edu

Expected Skills

- This class will use a fair amount of **math**
  - Probability and Statistics
  - Some 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
- Early assignments will be mostly **math**, later will be **coding**

Course Overview

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

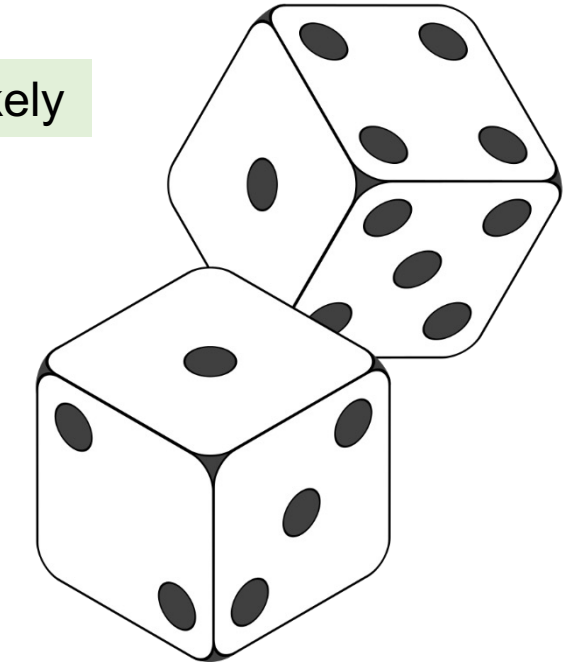
Probability and Statistics	Data Handling and Visualization	Machine Learning
Random events / variables, distributions / densities, moments, descriptive stats, estimation	Reading & cleaning, transformation & preprocessing, visualization	Predictive models, supervised learning, unsupervised learning, model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

↑ more on this in CSC 480/580

Probability and Statistics

Suppose we roll <u>two fair dice</u>…

fair die: each side is equally likely



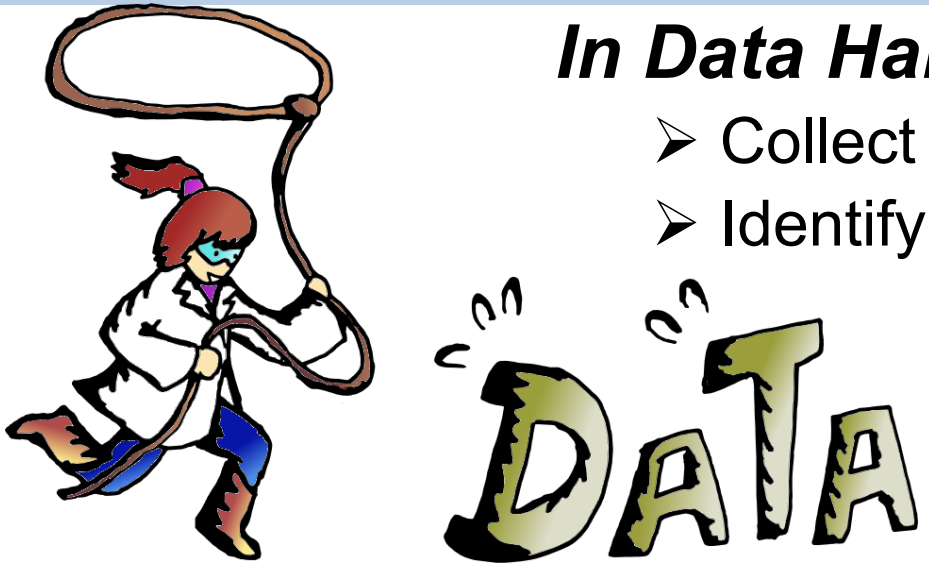
- What are the possible outcomes?
- What is the *probability* of rolling **even** numbers?

*… 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



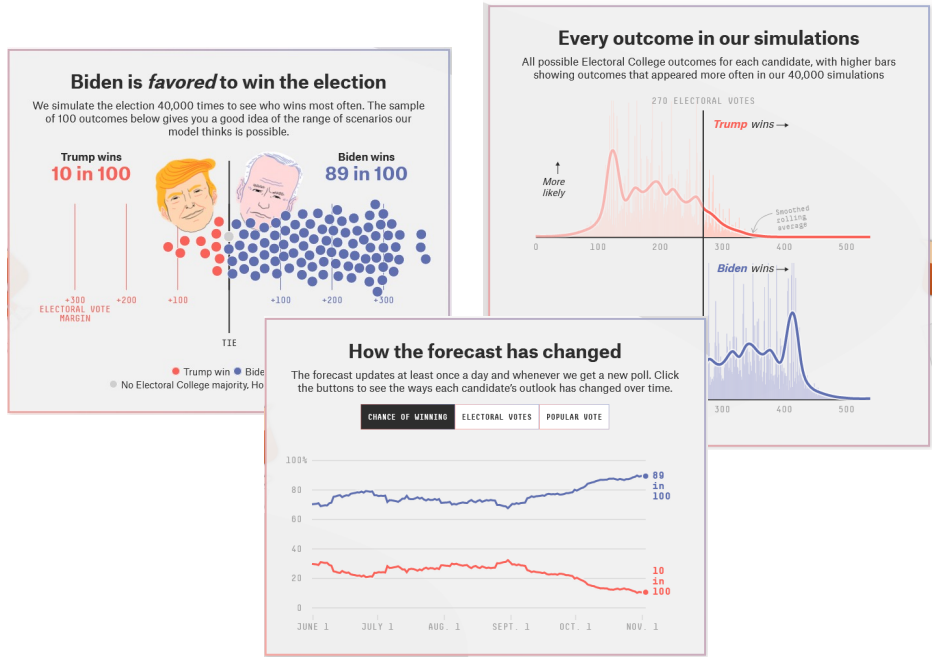
[Image Source: Code A Star]

In Data Handling we will learn to…

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

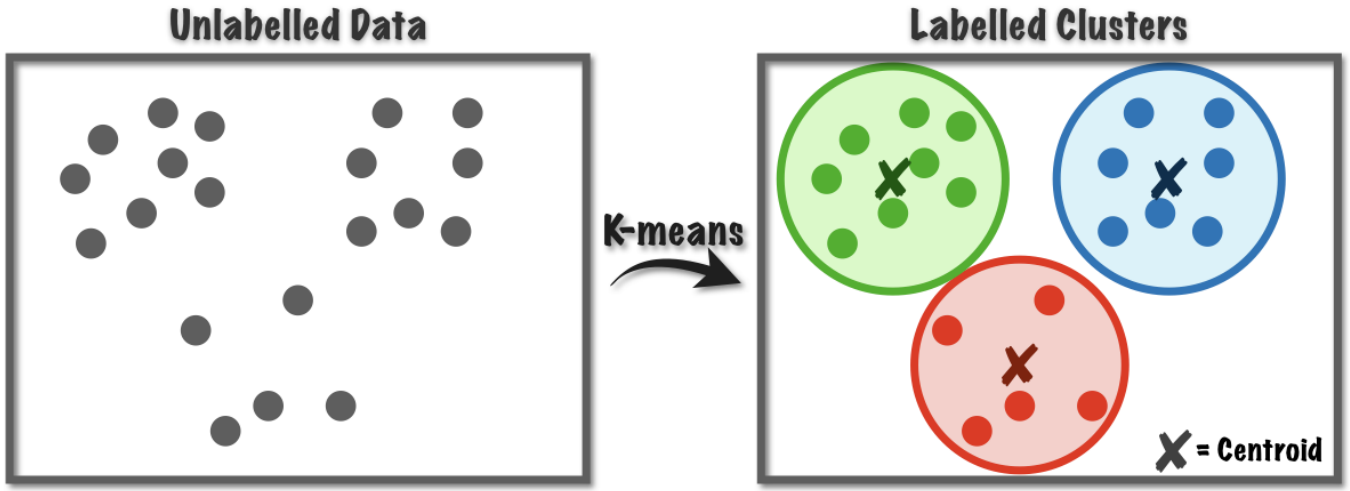
In Data Visualization we will learn…

- 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 Science]

Assignments / Exams / Grading

7 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-HW3 individually or 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 HW with the lowest score will be dropped

Grading Breakdown

- Assignments: 36% (6% each)
- Midterm: 20%
- Project: 14%
- Final Exam: 20%
- Participation: 10%

First assignment out next Tuesday

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 Kaggle competition.
- A guided project. You will answer given questions, including some open questions.
- You will get a change to try out various ML algorithms and get high accuracy.
- For top 10%, extra score (+2%).

Communication

- Announcements will be made via Piazza

- Homework submission: **gradescope** (see course website for the link)
  - Make sure your gradescope email address is the same as your D2L's

- **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.

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, once a week.
- 1hr by each TA, once a week.
- The final office hour schedule will be announced at the end of this week.
- If you have a conflict with the schedule, let me know (Piazza)

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

Full Course Schedule (Tentative)

Tentatitve; We will constantly update the schedule page

Dates	Topics	Homework
Aug 22 Aug 24	Course Overview Probability 1	
Aug 29 Aug 31	Probability 2 Probability 3	HW1 Out
Sep 7	Probability 4	
Sep 12 Sep 14	Probability 5 Probability 6	HW2 Out
Sep 19 Sep 21	Statistics 1 Statistics 2	HW3 Out
Sep 26 Sep 28	Statistics 3 Statistics 4	
Oct 3 Oct 5	Data processing and visualization 1 Midterm review	HW4 Out
Oct 10 Oct 12	Data processing and visualization 2 MIDTERM	
Oct 17 Oct 19	Data processing and visualization 3 Basics of predictive modeling and classification 1	HW5 Out
Oct 24 Oct 26	Basics of predictive modeling and classification 2 Basics of predictive modeling and classification 3	
Oct 31 Nov 2	Linear models 1 Linear models 2	HW6 Out
Nov 7 Nov 9	Linear models 3 Nonlinear models 1	
Nov 14 Nov 16	Nonlinear models 2 Nonlinear models 3	HW7 Out Final project Out
Nov 21	Clustering 1	
Nov 28 Nov 30	Clustering 2 Course wrap-up 1	
Dec 5	Course wrap-up 2	
Dec 13	FINAL EXAM	

Important Dates

- Sep 1: last date to self-withdraw without a ‘W’
- Oct 29: last date to withdraw
  - $\geq$ 40% of your total grade will be available by then.
- Nov 16: final project out
- Dec 8: final project due
- Dec 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 week 1 page)

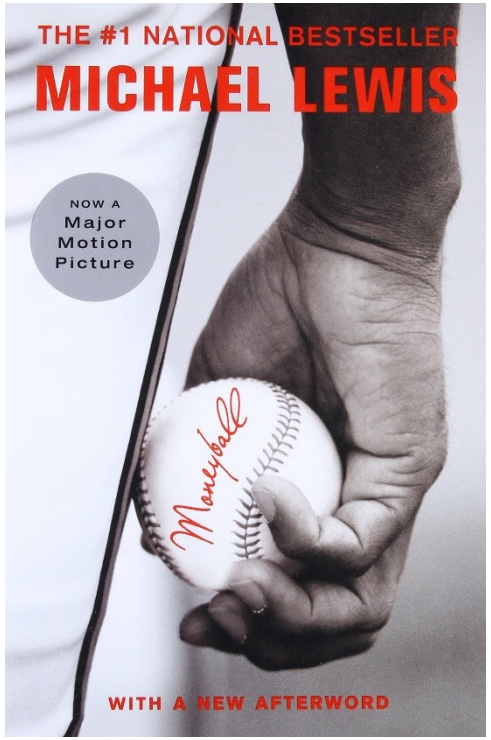
- 'Probability and statistics cookbook' is a good cheat sheet. Download it from http://statistics.zone/

Thank you

Moneyball

Problem *How to assemble the best baseball team with a small budget?*

- Story about the Oakland Athletics baseball team and its general manager **Billy Beane** for 2002 Major League Baseball (MLB) draft
- Traditional team building relies on *scouts* – but they are often biased and flawed.
- **SABRmetrics:** Data-driven and evidence-based approach to player quality evaluation
- *On-base %* and *Slugging %* are good indicators of offensive success
- Players with these “features” are cheaper compared to traditional statistics (stolen bases, runs batted in, batting average)



On-base %: how frequently a batter reaches base
Slugging %: the total number of bases a player records per at-bat

Moneyball: Impact

- In 2002 the Oakland Athletics ($44M budget) were competitive to the New York Yankees ($125M budget)
- Toronto Blue Jays hired full-time sabermetric analysts
- 2020 season “masters of Moneyball” Tampa Bay Rays reached world series with the 3rd lowest salary of all MLB
- In 2019 Liverpool Football/Soccer adopted this approach to nearly win the title (they lost to Manchester)
- Brad Pitt got a paycheck out of it for the movie (7.6/10 IMDB)…