INTRODUCTION TO DATA SCIENCE

JOHN P DICKERSON

Lecture #1 – 09/01/2021

CMSC320
Tuesdays & Thursdays
5:00pm – 6:15pm

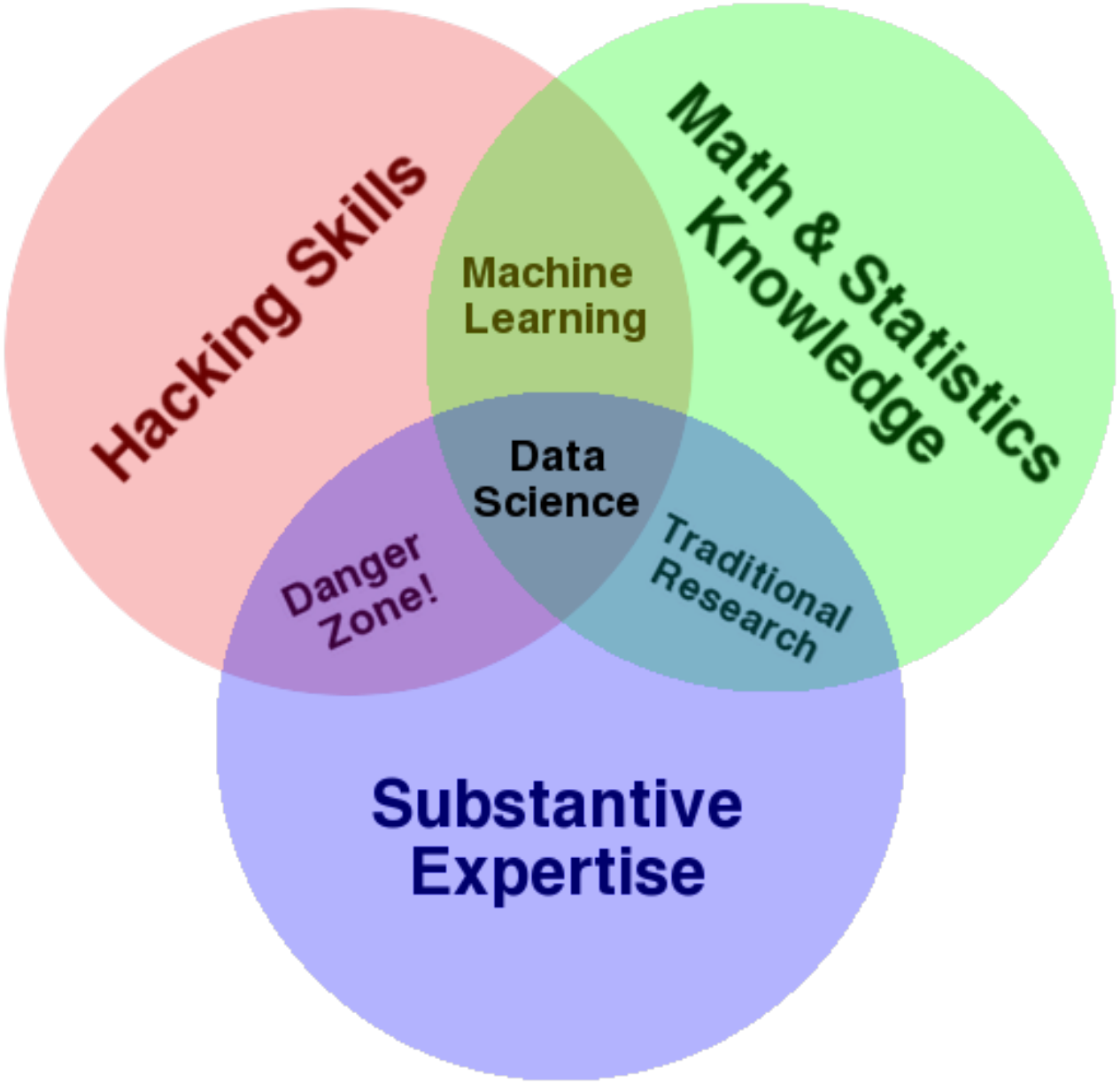
https://cmsc320.github.io/

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

INTRODUCTION TO ??????????

Data science is the application of computational and statistical techniques to address or gain [managerial or scientific] insight into some problem in the real world.

Zico Kolter
Machine Learning Prof, CMU
Chief Scientist of AI Research, Bosch



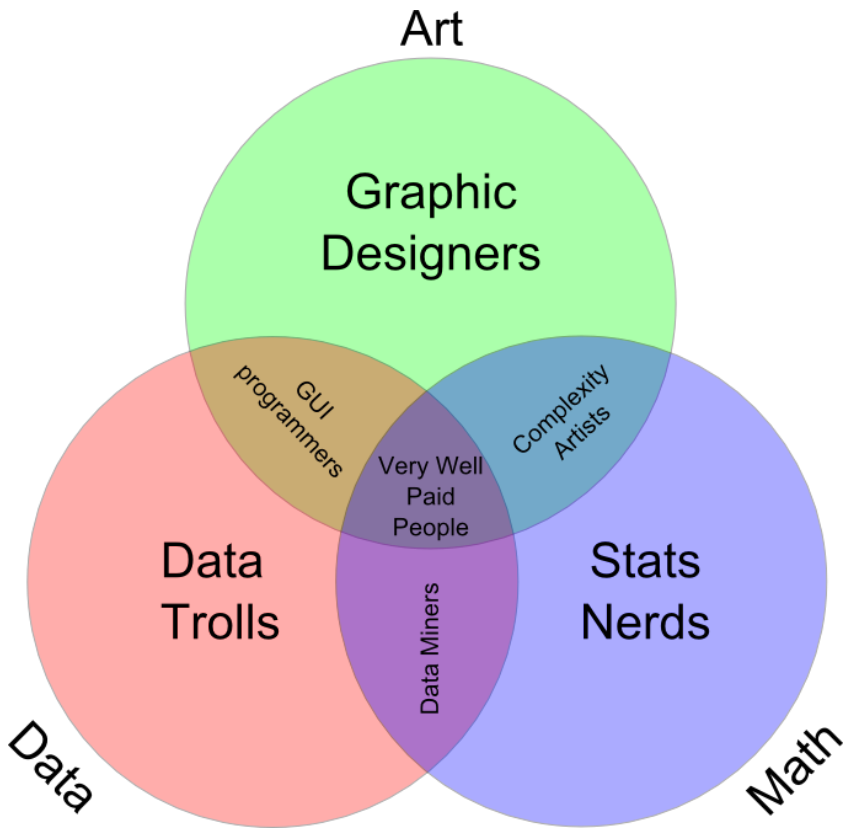
Drew Conway
CEO, Alluvium (analytics company)

MANY DEFINITIONS

Broad: necessarily **larger** than a single discipline

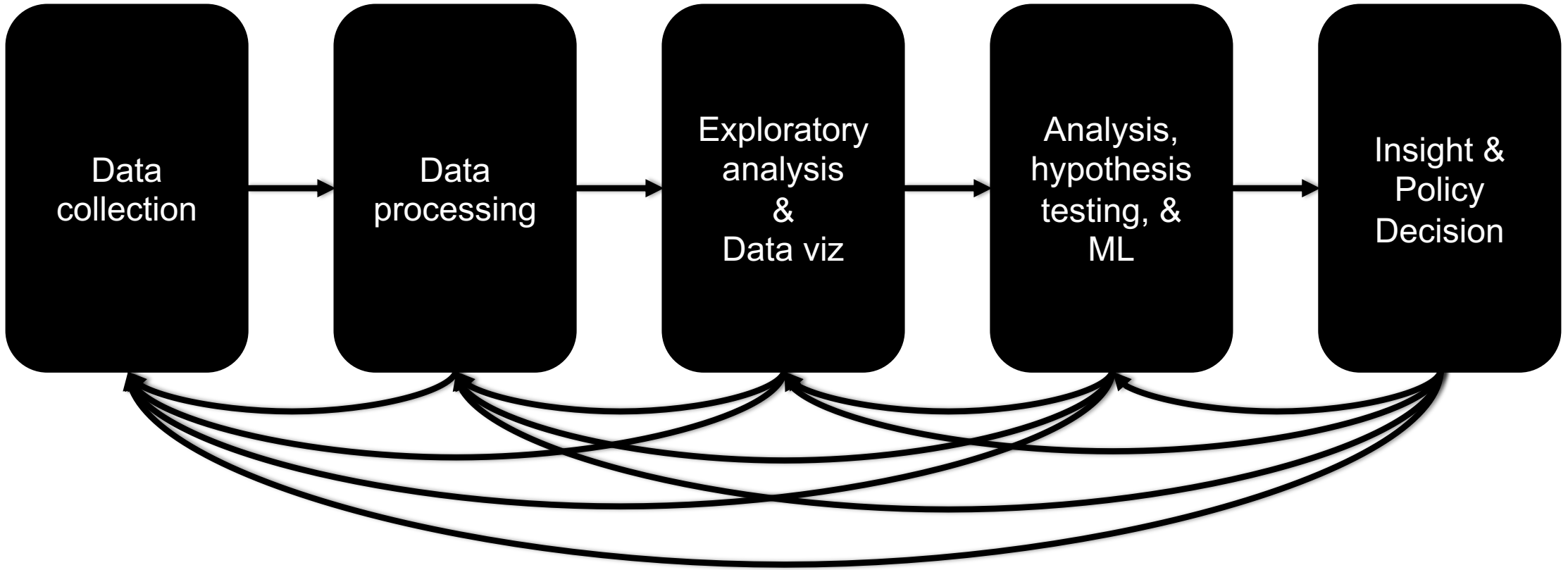
Interdisciplinary: statistics, computer science, operations research, statistical and machine learning, data warehousing, visualization, mathematics, information science, …

Insight-focused: grounded in the desire to find insights in data and leverage them to inform decision making



Tuomas Carsey, UNC

THE DATA LIFECYCLE



"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades …"

Hal Varian
Chief Economist at Google

THIS COURSE

You'll learn to take data:

- **Process it**
- **Visualize it**
- **Understand it**
- **Communicate it**
- **Extract value from it**

Hal Varian

Info: https://cmsc320.github.io/

Piazza: piazza.com/umd/fall2021/cmsc320

ELMS: (everyone should be registered automatically)

Panopto recordings available via ELMS / as normal

COURSE STAFF

Instructor: John Dickerson (just call me “John”)

- **Office location: Iribe 4128** (elevators to floor 4, take a right, then a right, then go straight and look at the numbers until you see the office that says “4128” or “John Dickerson” or has John sitting in it)
- **Office hours:** TBD, will be posted on Piazza/ELMS/lecture notes
- **More info:** http://jpdickerson.com/

Teaching Assistants (and potential research connections!):

- **Office hours:** TBD, will be posted on Piazza/ELMS (will cover MTWThF)

- Hirunima Jayasekara
- Kamala Varma
- MG Hirsch
- Alexander Gao
- Tobias Janssen
- Fuxiao Liu
- Neel Jain
- Sazan Mahbub

PREREQUISITE KNOWLEDGE

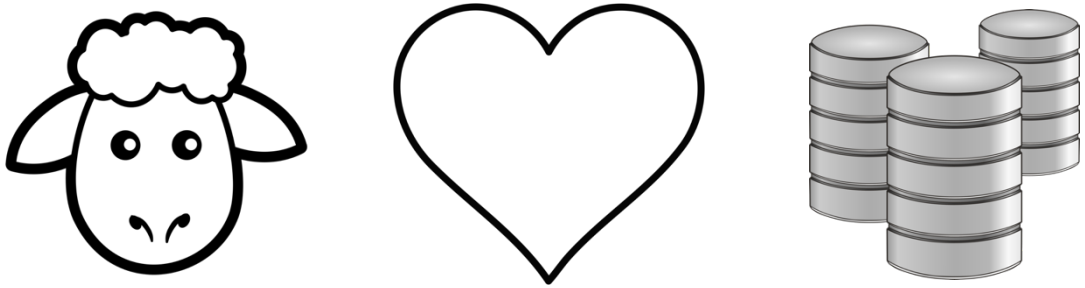
Aimed at CMSC undergrads – but likely accessible to others with programming experience and mathematical maturity.

We do not assume:

- Experience with Python, NumPy, pandas, scikit-learn, PyTorch, matplotlib, etc …
- Deep statistics or any ML knowledge
- Database or distributed systems knowledge

We do assume:

- You want to be here!



WHO AM I?

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND 2004 – 2008

Carnegie Mellon University
School of Computer Science 2010 – 2016

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND 2016 – ∞

UMIACS HCIL CENTER FOR MACHINE LEARNING
APPLIED MATHEMATICS & STATISTICS, AND SCIENTIFIC COMPUTATION PROGRAM

- facebook research 2018 – 2020
- { Advisory roles in UMD alumni's startups } 2017 – now Arthur govshop
- { Advisory role to a US regulatory agency } Late 2019 - 2021

UNOS DONATE LIFE®
UNITED NETWORK FOR ORGAN SHARING

OPTIMIZED MARKETS

NATIONAL SECURITY AGENCY UNITED STATES OF AMERICA

IBM®

THE DIAN FOSSEY GORILLA FUND INTERNATIONAL

WHO ARE YOU?

2nd-year
3rd-year
4th-year +

STAT400?
CMSC422?
CMSC424?

python™

NumPy
NetworkX
Network Analysis in Python
scikit learn
PyTorch
TensorFlow

Register on Piazza: piazza.com/umd/fall2021/cmsc320

(TENTATIVE) COURSE STRUCTURE

First 4 lectures: intro & primers in the Python data science stack

Next 6 lectures: data collection & management

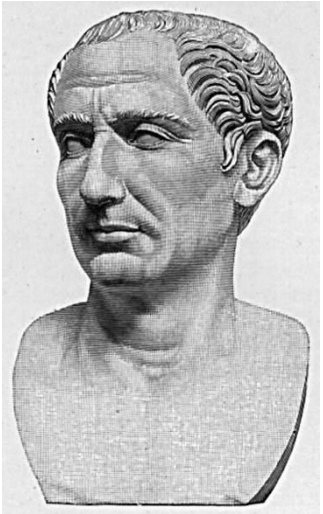
- Best practices, data wrangling, exploratory analysis, ethics, debugging, visualization, etc …

Next 9 lectures: statistical modeling & ML

- Statistical learning, regression, classification, cross-validation, model evaluation, hypothesis testing, etc …
- (Midterm #1 toward the beginning of this block)

Final 8 lectures: advanced topics

- Dimensionality reduction, distributed learning, big data, distributed computation
- (Midterm #2 toward the beginning of this block)



Ambitious …

GRADE #1: MINI-PROJECTS

Students will complete four mini-project assignments:

- **Case studies** meant to mimic what you, a future data scientist, will see in industry. They should be fun ☺.

The rules:

- Allowed: small group **discussions**
- Required: individual **programming & writing**
- Never allowed: public posting of solutions

Deliverable:

- Turn in an .ipynb of a Jupyter notebook on ELMS

GRADE #2: READING HOMEWORKS

We will post (bi)weekly reading assignments. Mix of:

- Blog posts
- Academic articles
- News articles

Weekly quiz to be taken on ELMS covering the readings

Individual quiz grades are pass/fail:

- At least 60% correct → Pass
- Less than 60% correct → Fail

Must take at least ten of these quizzes over the semester

GRADE #3: TWO MIDTERMS

You know what these are.

Will cover roughly the first 1/3, and then second 1/3, of class:

- Qualitative (more)
- Quantitative (less)

Currently scheduled for mid-October and mid-November

- May change, but will nail down exact dates soon.

GRADE #4: MINI-TUTORIAL

In lieu of a final exam, you'll create a mini-tutorial that:

- Identifies a raw data source
- Processes and stores that data
- Performs exploratory data analysis & visualization
- Derives insight(s) using statistics and ML
- Communicates those insights as actionable text

Individual or group project – see rubric for exact details, more information.

Will be hosted publicly online (GitHub Pages) and will strengthen your portfolio.

READY-MADE DATASET REPOSITORIES

https://www.data.gov/

- US-centric agriculture, climate, education, energy, finance, health, manufacturing data, …

https://cloud.google.com/bigquery/public-data/

- BigQuery (Google Cloud) public datasets (bikeshare, GitHub, Hacker News, Form 990 non-profits, NOAA, …)

https://www.kaggle.com/datasets

- Microsoft-owned, various (Billboard Top 100 lyrics, credit card fraud, crime in Chicago, global terrorism, world happiness, …)

https://aws.amazon.com/public-datasets/

- AWS-hosted, various (NASA, a bunch of genome stuff, Google Books n-grams, Multimedia Commons, …)

NEW DATASET IDEAS

Fraternal Order of Police vs Black Lives Matter

Linking finance data to ${anything_else}

Something having to do with Pokémon statistics?

Look through http://www.alexa.com/topsites and scrape something interesting!

University of Maryland-related, or College Park-related, stuff

- Check out http://umd.io/ – open source project; maybe your data collection and cleaning scripts can be added to this!

Honestly, pretty much anything! Just document everything.

Reproducibility!

FINAL TUTORIAL

Deliverable: URL of your own GitHub Pages site hosting an .ipynb/.html export of your final tutorial

- https://pages.github.com/ – make a GitHub account, too!
- https://github.com/blog/1995-github-jupyter-notebooks-3

The project itself:

- ~1500+ words of Markdown prose
- ~150+ lines of Python
- Should be viewable as a **static webpage** – that is, if I (or anyone else) opens the link up, everything should render and I shouldn't have to run any cells to generate output

Rubric with more details & example tutorials linked to from website!!

FINAL TUTORIAL RUBRIC

The TAs and I will grade on a scale of 1-10:

Motivation: Does the tutorial make the reader believe the topic is important (a) in general and (b) with respect to data science?

Understanding: After reading the tutorial, does the reader understand the topic?

Further resources: Does the tutorial "call out" to other resources that would help the reader understand basic concepts, deep dive, related work, etc?

Prose: Does the prose in the Markdown portion of the .ipynb add to the reader's understanding of the tutorial?

Code: Does the code help solidify understanding, is it well documented, and does it include helpful examples?

Subjective Evaluation: If somebody linked to this tutorial from Hacker News, would people actually read the whole thing?

Thanks to: Zico Kolter

CLASS PARTICIPATION

Please please please please please do the required reading, if available, before (virtually) attending class!

Earn brownie points via:

- Lecture participation
- Piazza participation
- Regular participation at office hours

Aim to ask/answer a question at least once every two weeks; or attend office hours at least once a month

GRADE BREAKDOWN

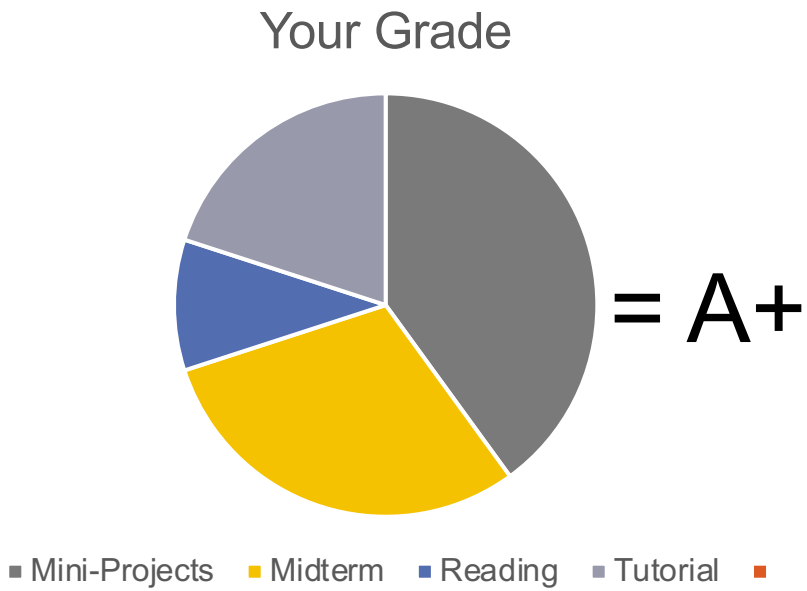
40% mini-projects:

- **There are 4 of them**
- **Equal weighting @ 10%**

10% reading homeworks (aka weekly quizzes)

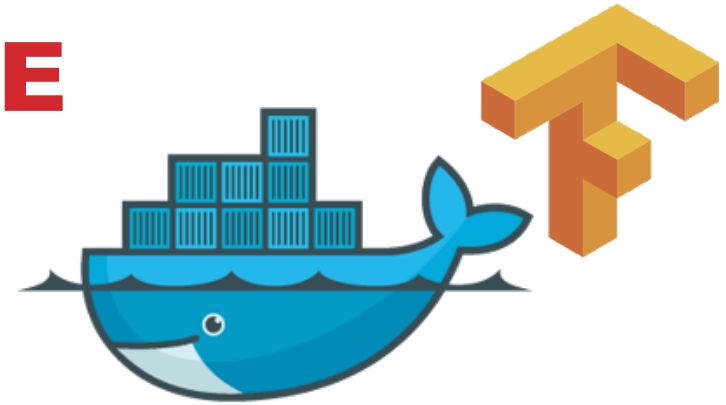
30% two midterms (15% each)

20% final tutorial



SOME TECHNOLOGIES WE WILL USE (MOSTLY)

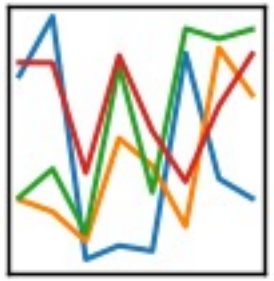
python™



docker

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



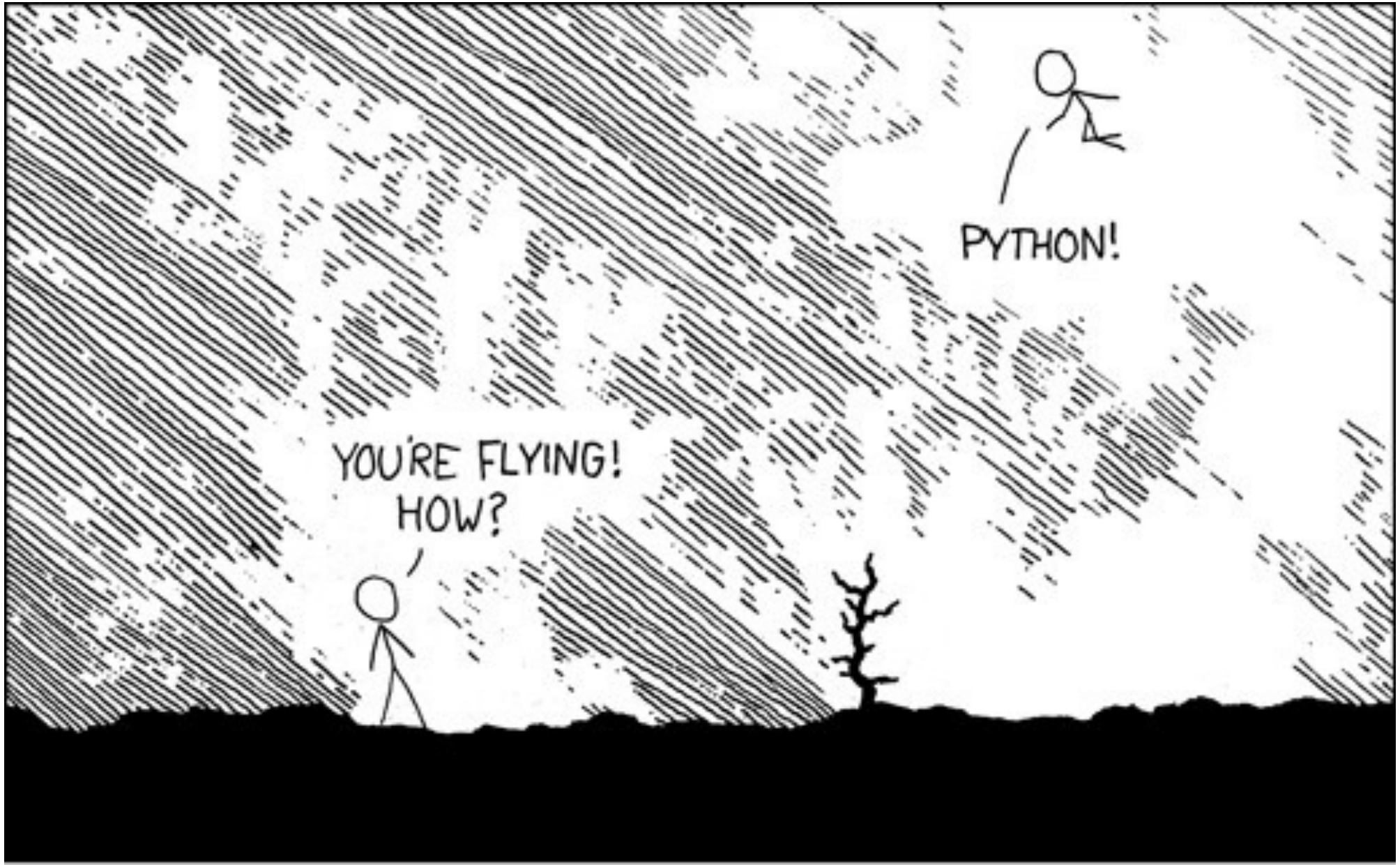
PyTorch

NetworkX
Network Analysis in Python

Spark

scikit learn

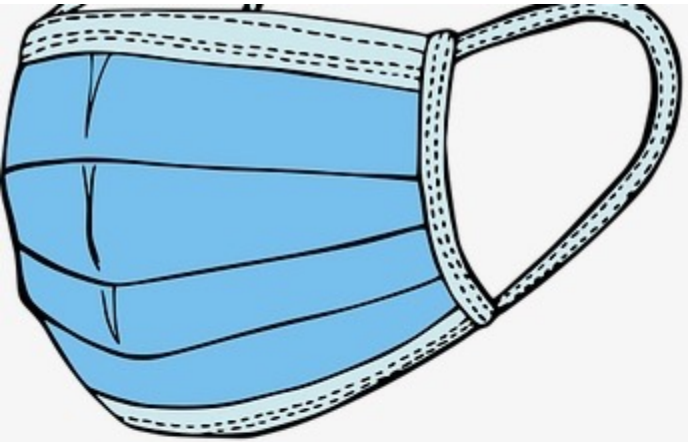
NumPy

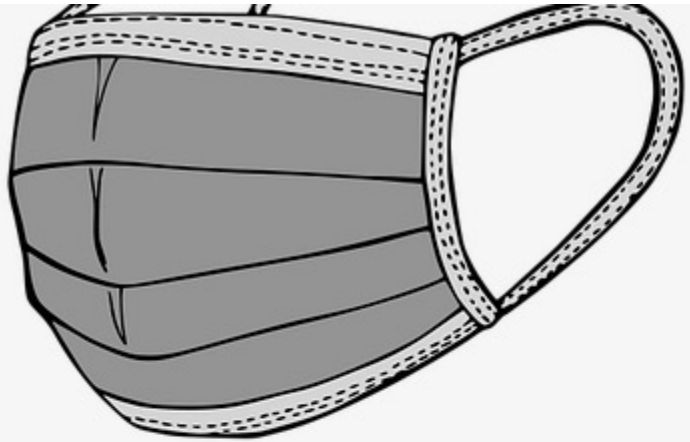


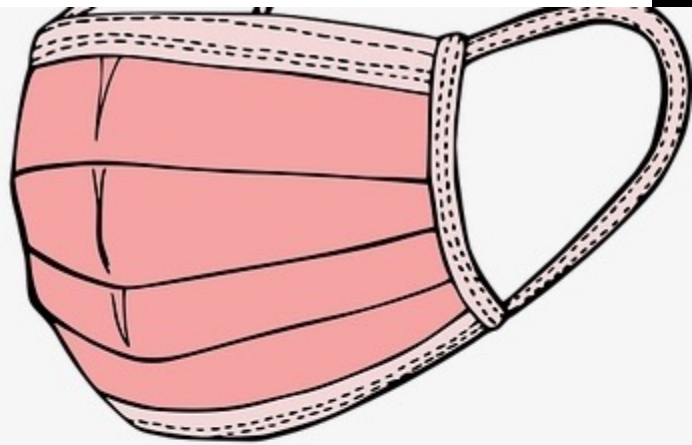
(Don't tell CMSC330 …)

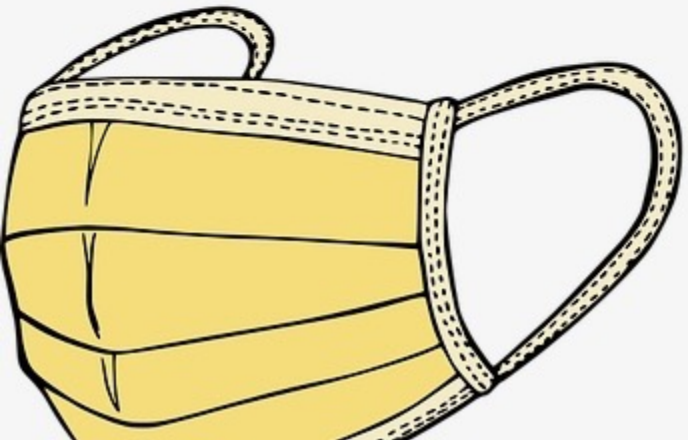
IMPORTANT WALLS OF TEXT

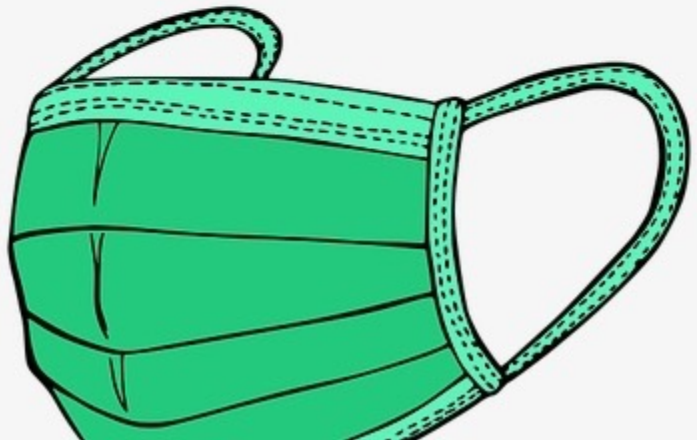
PLEASE WHERE A MASK (IT'S POLICY! DO IT!)

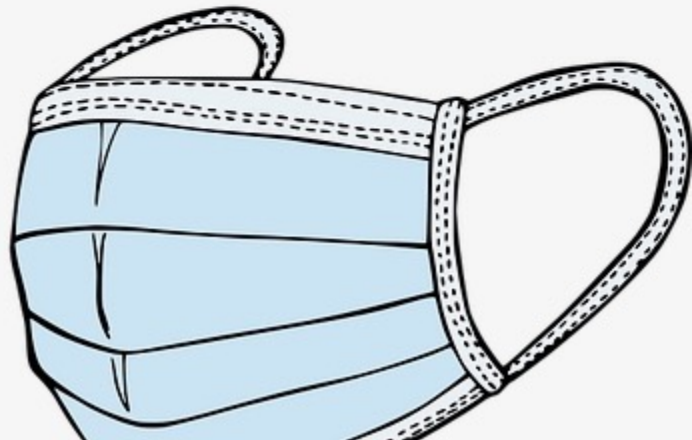












ANTI-HARASSMENT

(Adapted from ACM SIGCOMM's policies)

Common Sense!

The open exchange of ideas and the freedom of thought and expression are central to our aims and goals. These require an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity. For these reasons, we are dedicated to providing a harassment-free experience for participants in (and out) of this class.

Harassment is unwelcome or hostile behavior, including speech that intimidates, creates discomfort, or interferes with a person's participation or opportunity for participation, in a conference, event or program.

ACADEMIC INTEGRITY

Common Sense!

(Text unironically stolen from Hal Daumé III)

Any assignment or exam that is handed in must be your own work (unless otherwise stated). However, talking with one another to understand the material better is strongly encouraged. Recognizing the distinction between cheating and cooperation is very important. If you copy someone else's solution, you are cheating. If you let someone else copy your solution, you are cheating (this includes *posting solutions online in a public place)*. If someone dictates a solution to you, you are cheating.

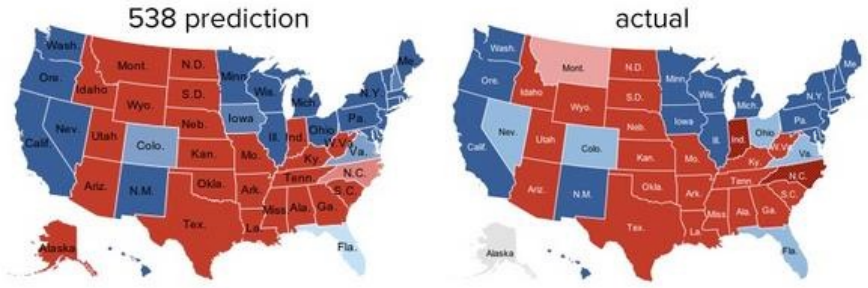
Everything you hand in must be in your own words, and based on your own understanding of the solution. If someone helps you understand the problem during a high-level discussion, you are not cheating. We strongly encourage students to help one another understand the material presented in class, in the book, and general issues relevant to the assignments. When taking an exam, you must work independently. Any collaboration during an exam will be considered cheating. Any student who is caught cheating will be given an F in the course and referred to the University Office of Student Conduct. Please don't take that chance – if you're having trouble understanding the material, please let me know and I will be more than happy to help.

(A FEW) DATA SCIENCE SUCCESS STORIES & CAUTIONARY TALES

POLLING: 2008 & 2012

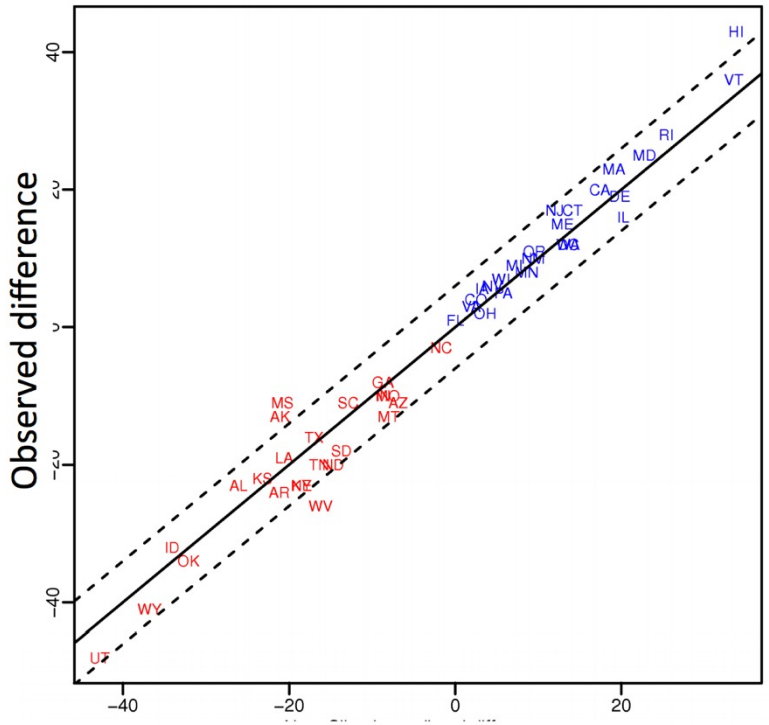
Nate Silver uses a simple idea – taking a principled approach to aggregating polling instead of relying on punditry – and:

- **Predicts 49/50 states in 2008**
- **Predicts 50/50 states in 2012**



- **(He is also a great case study in creating a brand.)**

https://hbr.org/2012/11/how-nate-silver-won-the-2012-p



Democrat (+) or Republican (-) in 2012

POLLING: 2016

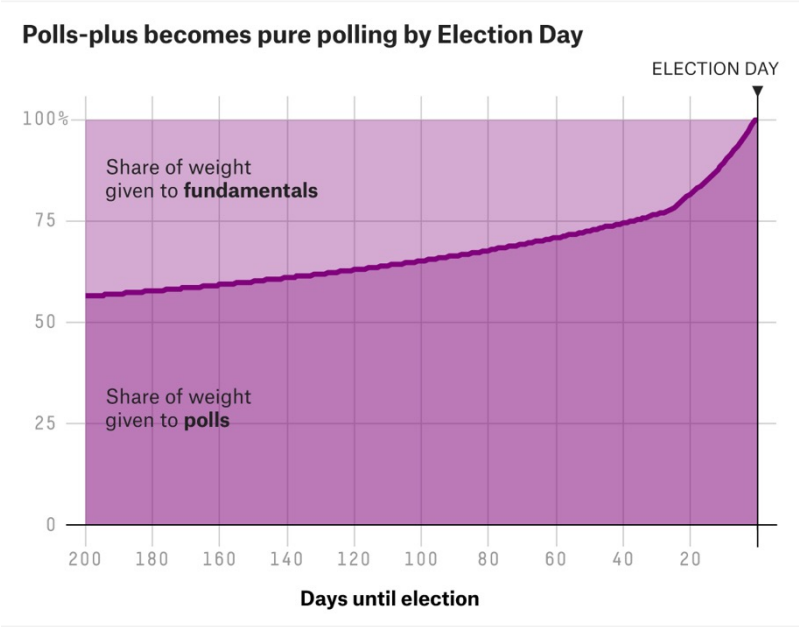
POLITICS

Nate Silver Is Unskewing Polls — All Of Them — In Trump's Direction

The vaunted 538 election forecaster is putting his thumb on the scales.

HuffPo: "He may end up being right, but he's just guessing. A "trend line adjustment" is merely political punditry dressed up as sophisticated mathematical modeling."

538: Offers quantitative reasoning for re-/under-weighting older polls, & changing as election approaches

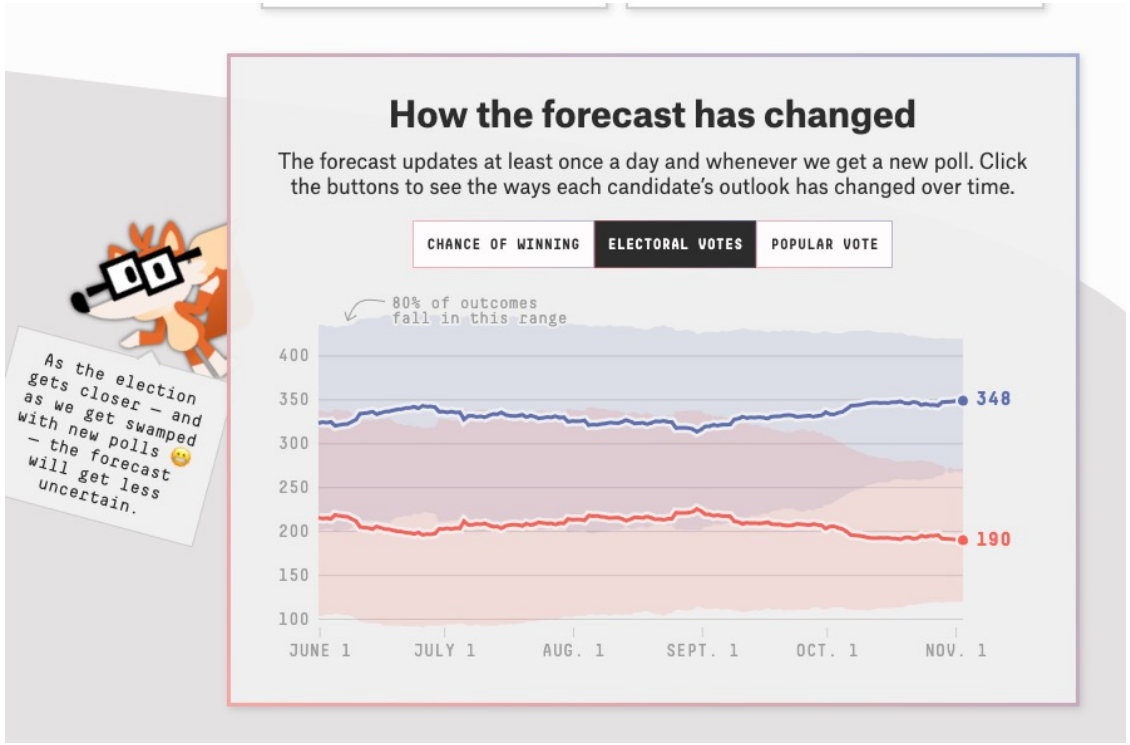


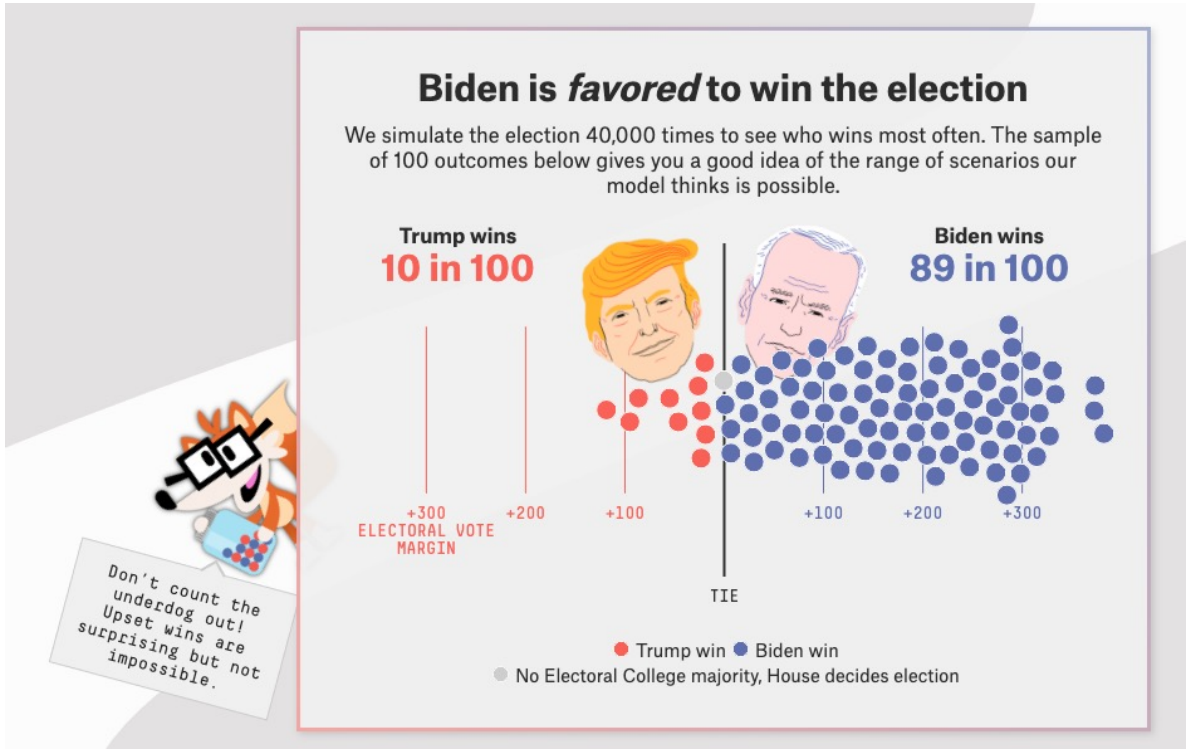
http://www.huffingtonpost.com/entry/nate-silver-election-forecast_us_581e1c33e4b0d9ce6fbc6f7f
https://fivethirtyeight.com/features/a-users-guide-to-fivethirtyeights-2016-general-election-forecast/

POLLING: 2020

Lessons learned: don't just communicate a binary prediction (X loses, Y wins):

- Communicate uncertainty over that prediction
- Communicate variance in the underlying model(s), modeling errors, etc





AD TARGETING

Pregnancy is an expensive & habit-forming time

- **Thus, valuable to consumer-facing firms**

2012:

- **Target identifies 25 products and subsets thereof that are commonly bought in early pregnancy**
- **Uses purchase history of patrons to predict pregnancy, targets advertising for post-natal products (cribs, etc)**
- **Good: increased revenue**
- **Bad: this can expose pregnancies – as famously happened in Minneapolis to a high schooler**

http://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2

AUTOMATED DECISIONS OF CONSEQUENCE

[Sweeney 2013, Miller 2015, Byrnes 2016, Rudin 2013, Barry-Jester et al. 2015]

textio

DOXA

zest finance

LendingClub

PREDPOL

gild

GJ GapJumpers

Kabbage

NORTHPOINTE
Advancing Justice. Embracing Community.

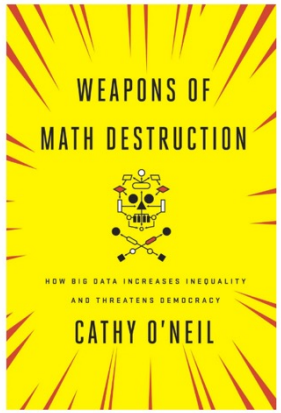
Hiring

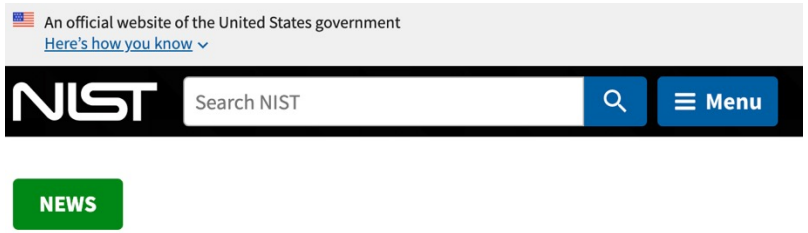
Lending

Policing/ sentencing

Search for minority names →
ads for DUI/arrest records

Female cookies →
less freq. shown professional job opening ads





NEWS

NIST Proposes Approach for Reducing Risk of Bias in Artificial Intelligence

Comments are sought on the publication, which is part of NIST's effort to develop trustworthy AI.

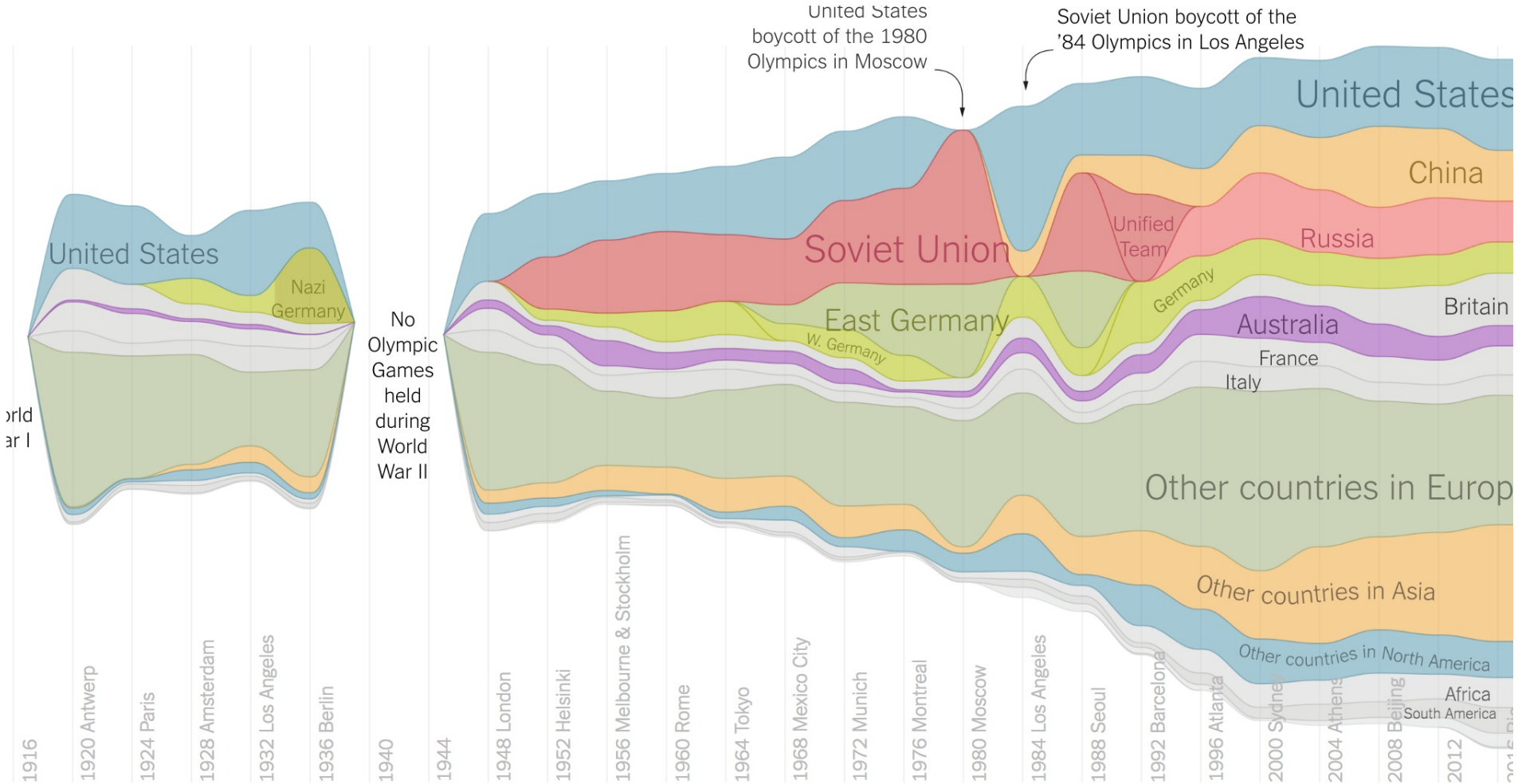
"… a lot remains unknown about how big data-driven decisions may or may not use factors that are proxies for race, sex, or other traits that U.S. laws generally prohibit from being used in a wide range of commercial decisions … What can be done to make sure these products and services–and the companies that use them treat consumers fairly and ethically?"

- FTC Commissioner Julie Brill [2015]

"**Hold yourself accountable – or be ready for the FTC to do it for you.** As we've noted, it's important to hold yourself accountable for your algorithm's performance. … keep in mind that if you don't hold yourself accountable, the FTC may do it for you."

- FTC official blog post, "Aiming for truth, fairness, and equity in your company's use of AI", written by Elisa Jillson [2021]

OLYMPIC MEDALS



https://www.nytimes.com/interactive/2016/08/08/sports/olympics/history-olympic-dominance-charts.html

NETFLIX PRIZE I

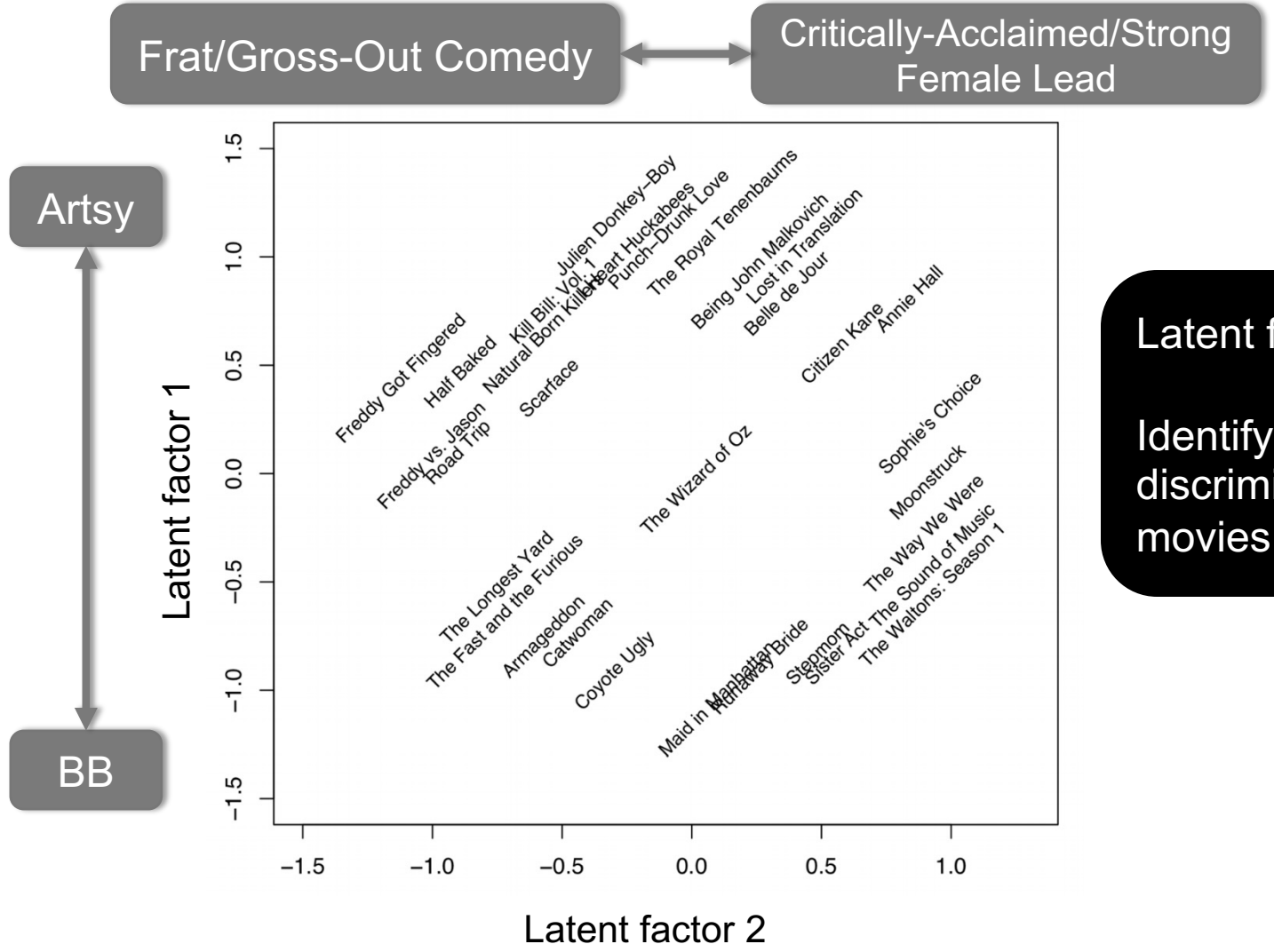
Recommender systems: predict a user's rating of an item

	Twilight	Wall-E	Twilight II	Furious 7
User 1	+1	-1	+1	?
User 2	+1	-1	?	?
User 3	-1	+1	-1	+1

Netflix Prize: $1MM to the first team that beats our in-house engine by 10%

- **Happened after about three years**
- **Model was never used by Netflix for a variety of reasons**
  - Out of date (DVDs vs streaming)
  - Too complicated / not interpretable

NETFLIX PRIZE II



Latent factors model:

Identify factors with max discrimination between movies

Image courtesy of Christopher Volinsky

NETFLIX PRIZE III

Netflix initially planned a follow-up competition

In 2007, UT Austin managed to deanonymize portions of the original released (anonymized) Netflix dataset:

- ??????????
- Matched rating against those made publicly on IMDb

Why could this be bad?

2009—2010, four Netflix users filed a class-action lawsuit against Netflix over

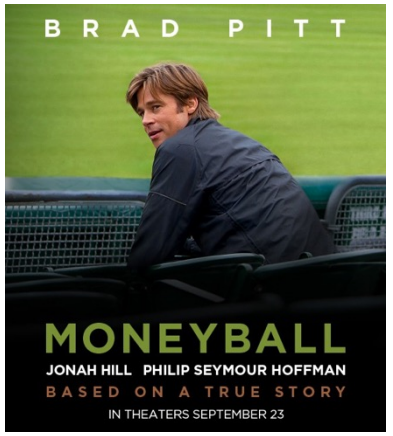
MONEYBALL

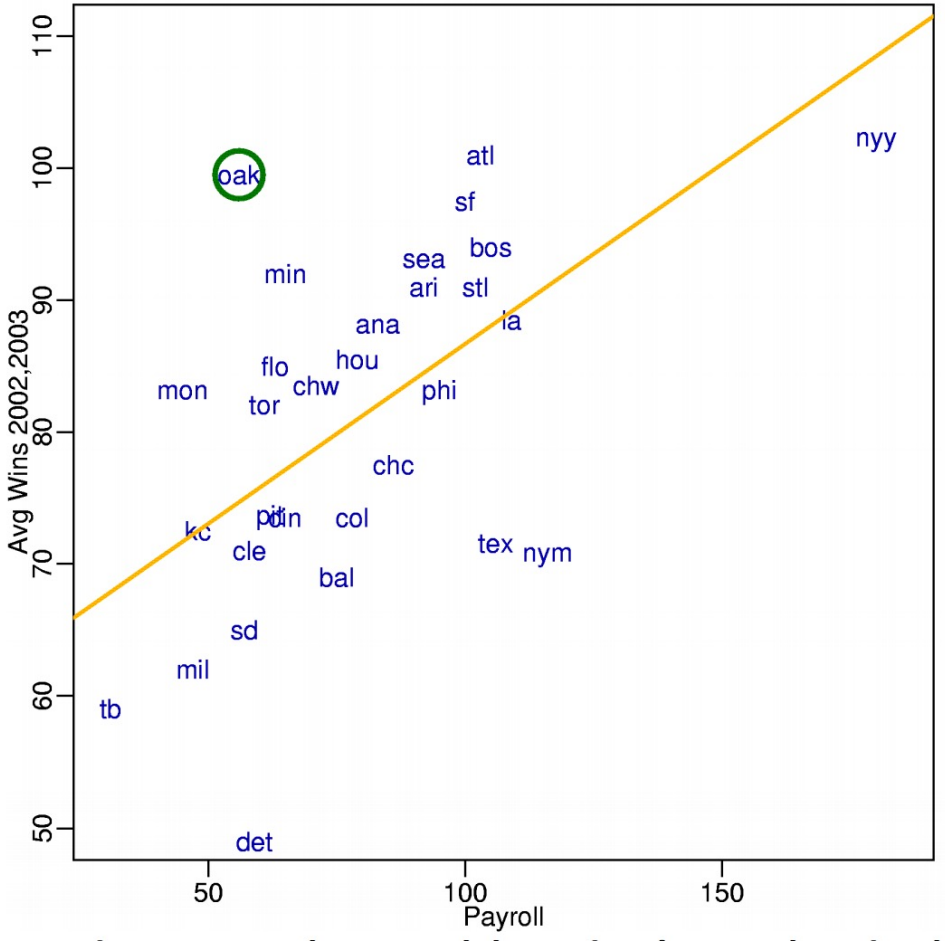
Baseball teams drafted rookie players primarily based on human scouts' opinions of their talents

Paul DePodesta, data scientist *du jour*, convinces the {bad, poor} Oakland Athletics to use a quantitative aka sabermetric approach to hiring

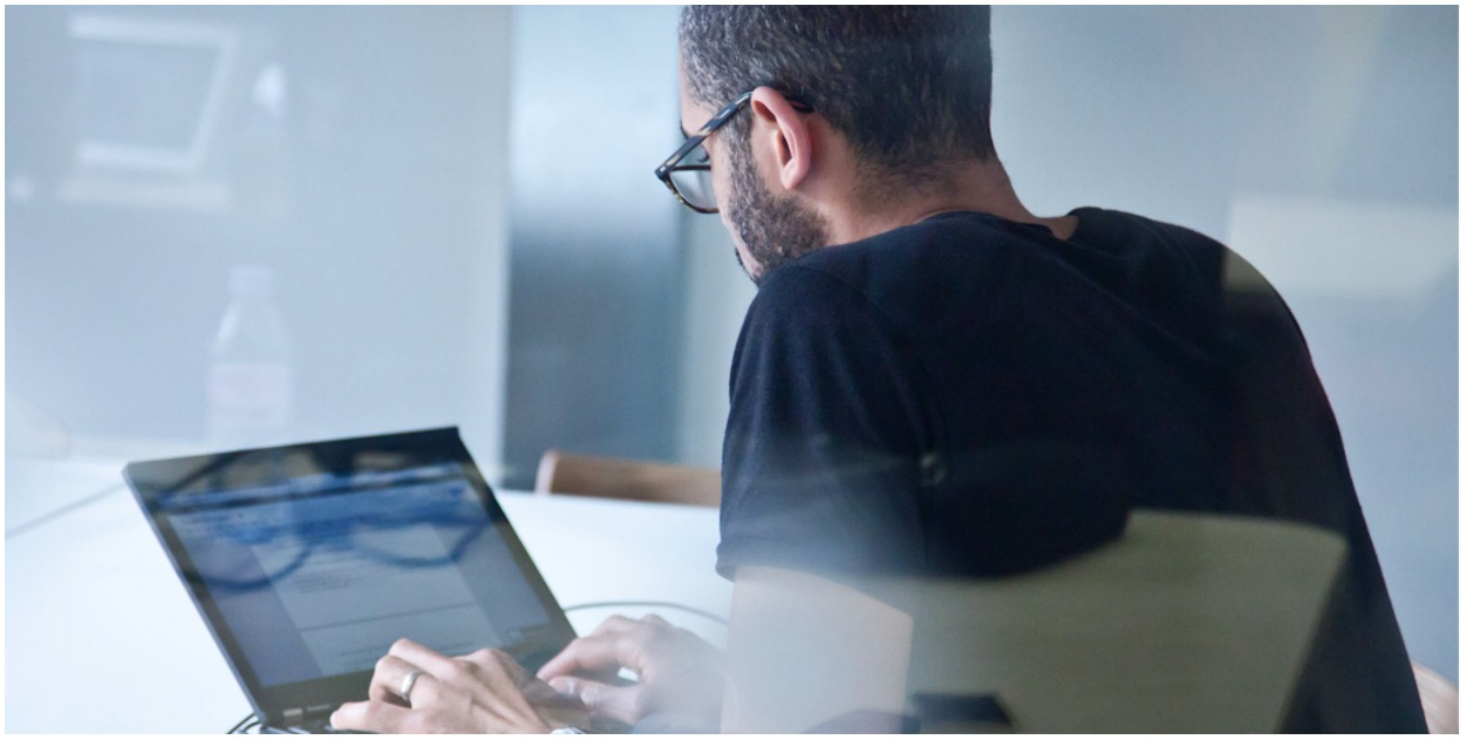
(Spoiler: Red Sox offer Brand a job, he says no, they take a sabermetric approach and win the World Series.)

(Spoiler #2: DePodesta is now Chief Strategy Officer for the Browns, and they extended his contract in 2021, so we'll see what happens!)





1. Data scientist



Shutterstock

Overall job score (out of 5.0): 4.8

Job satisfaction rating (out of 5.0): 4.4

Number of job openings: 4,184

Median base pay: $110,000

http://www.businessinsider.com/best-jobs-in-america-in-2017-2017-1/

WRAP-UP FOR TODAY

Register on Piazza using your UMD address:

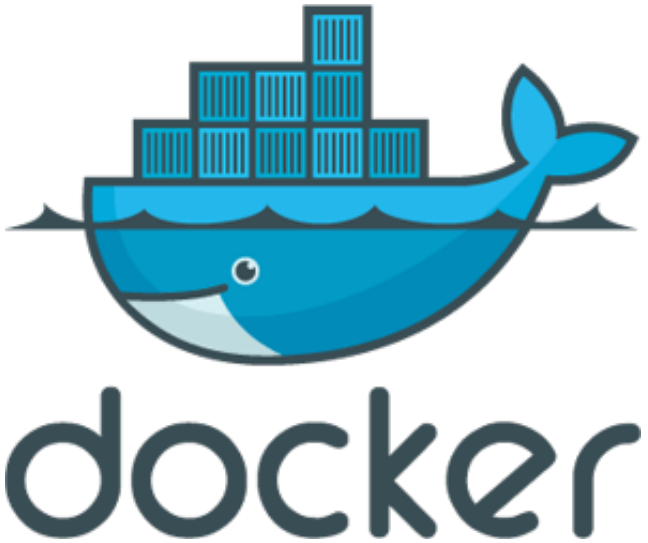
piazza.com/umd/fall2021/cmsc320

Please get in touch with me if you're unsure of whether or not you're at the right {programming, math} level for this course:

- My guess is that you are!
- This is a young class, so we're quite flexible

Tonight, read about Docker & Jupyter!

- Works on *nix, OSX, Windows
- https://www.docker.com/
- (We'll post a small project shortly.)



NEXT CLASS:

SCRAPING DATA WITH PYTHON