INTRODUCTION TO DATA SCIENCE

MOHAMMAD NAYEEM TELI

This class is being recorded

Lecture #1 – 09/01/2020

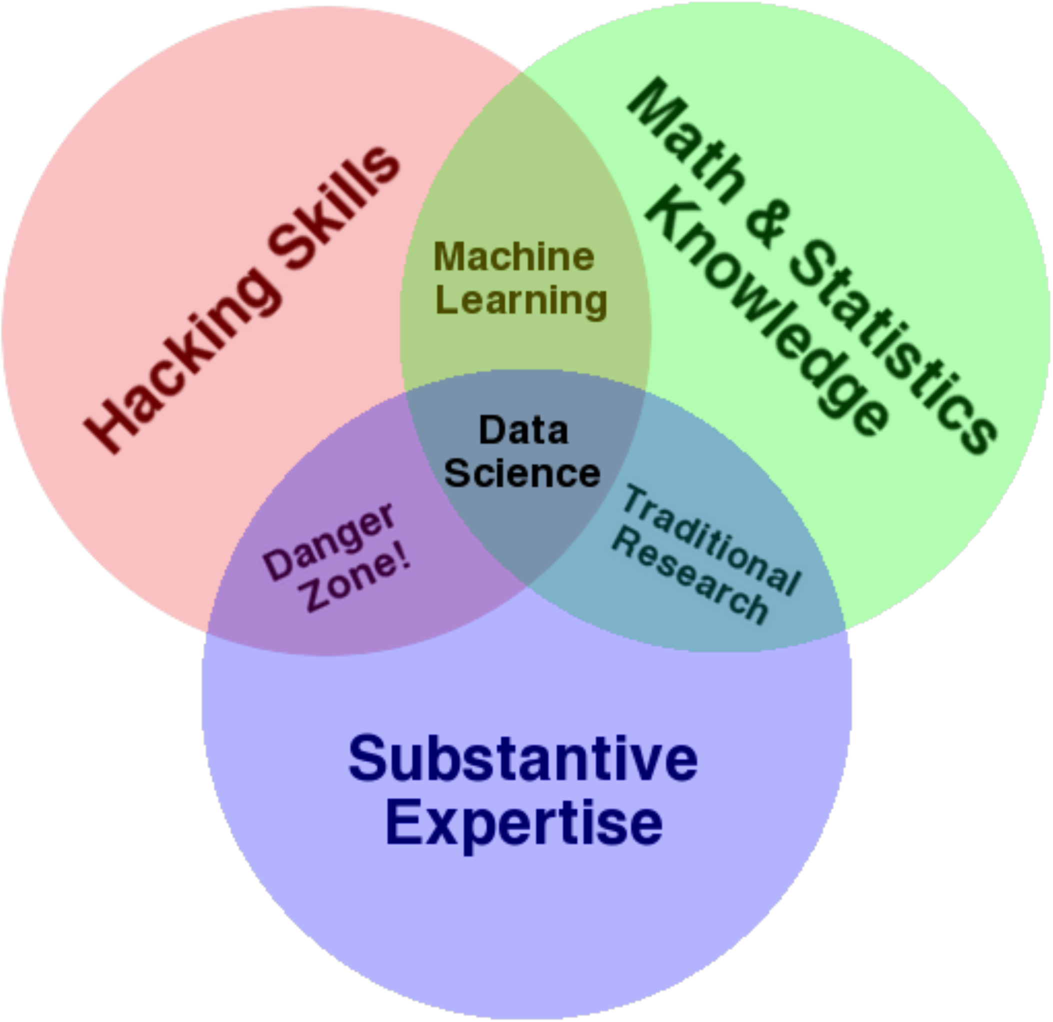
CMSC320
TuTh
11:00am – 12:15pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Thanks to John Dickerson for most of the slides

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



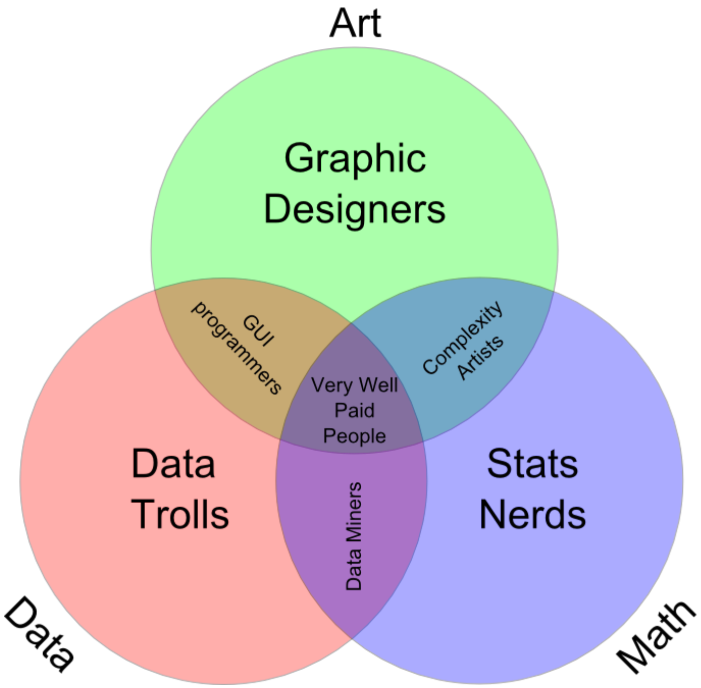
Drew Conway
CEO, Alluvium (analytics company)

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

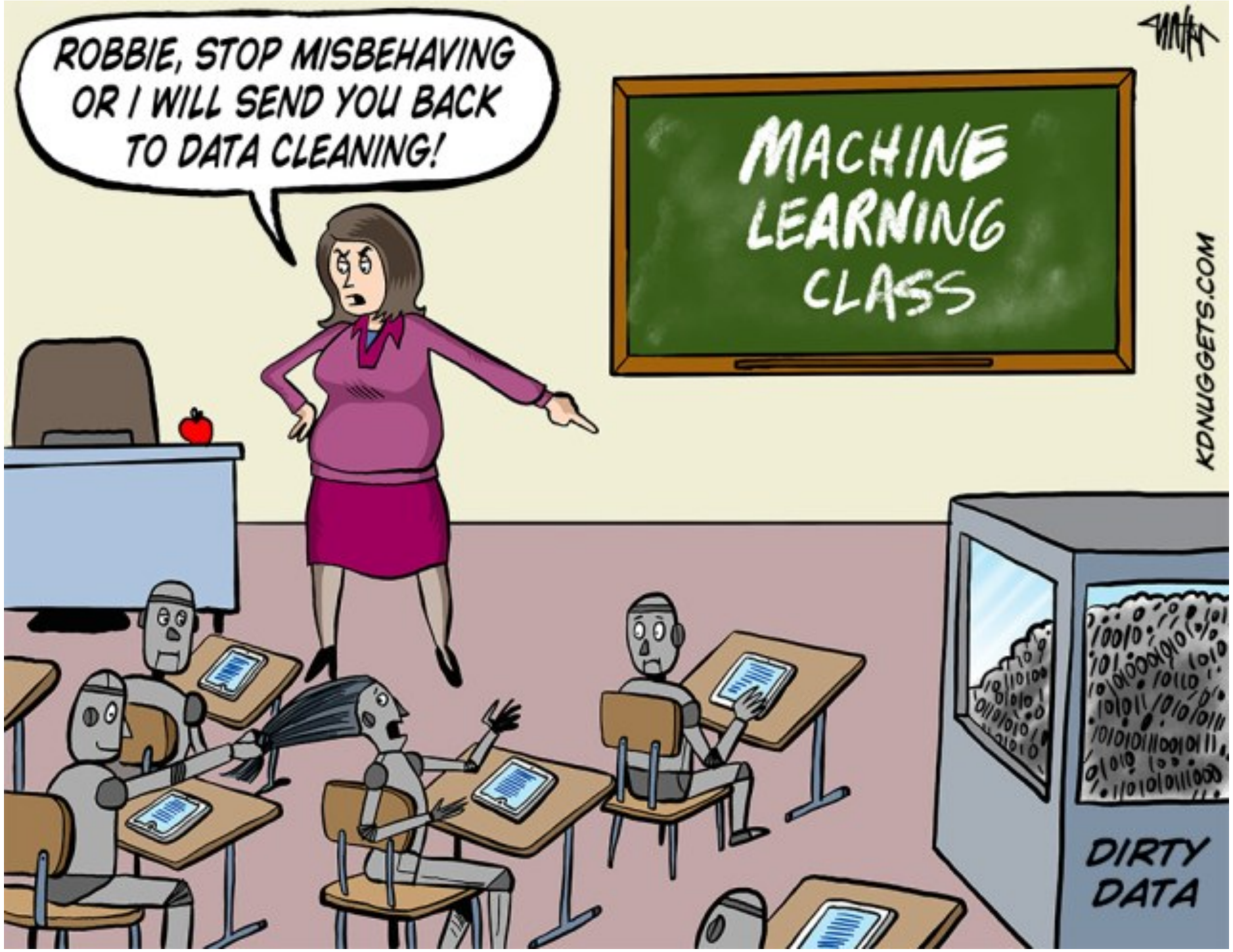
DATA SCIENCE LIFECYCLE

1. Collect Data

Remember Mr Pooter is not just a 'patient', he's an important source of valuable and readily marketable data!

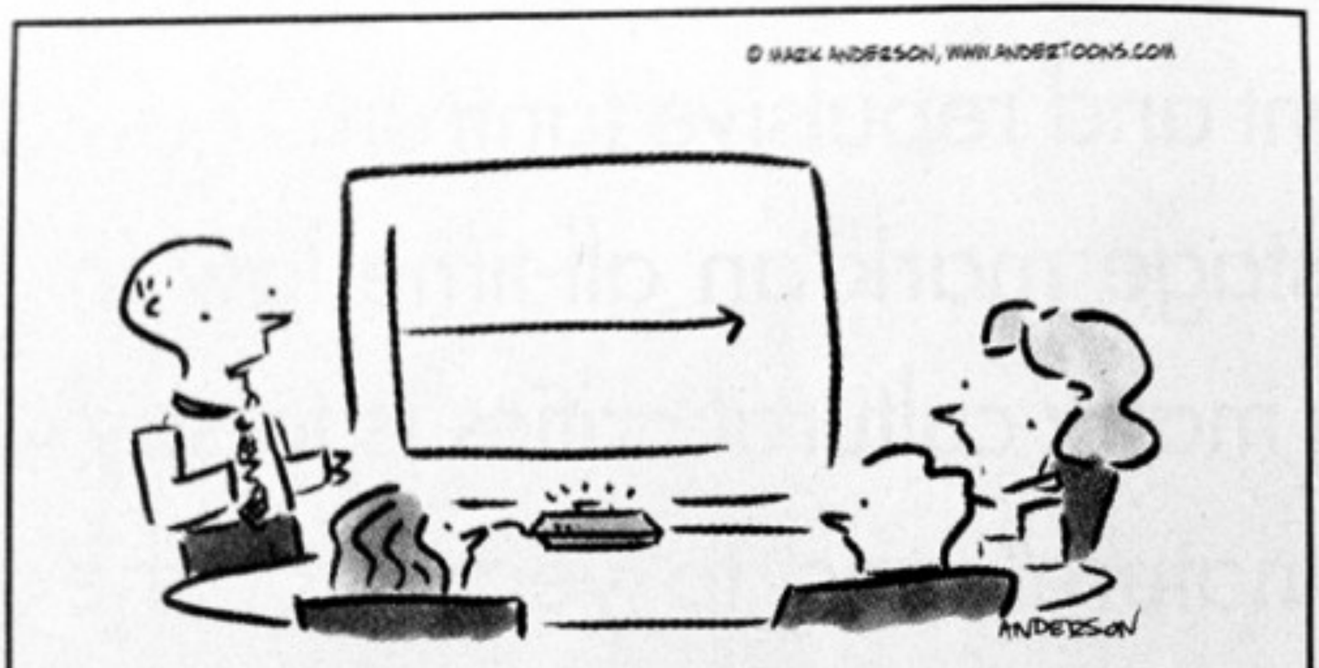
DATA SCIENCE LIFECYCLE

2. Scrub Data


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

DATA SCIENCE LIFECYCLE

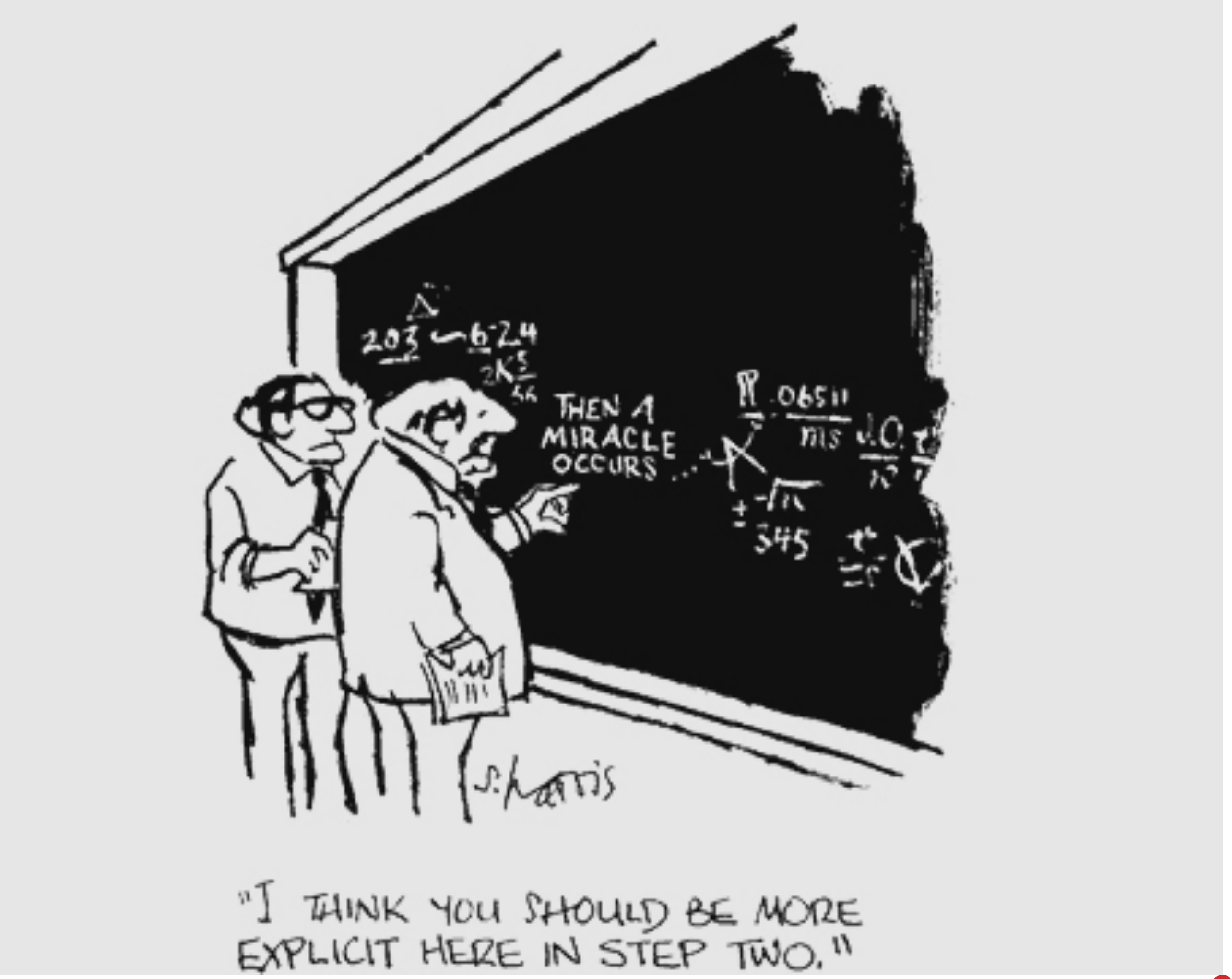
3. Explore Data



"After closer investigation, it's become clear that we need to enter more than one value."

DATA SCIENCE LIFECYCLE

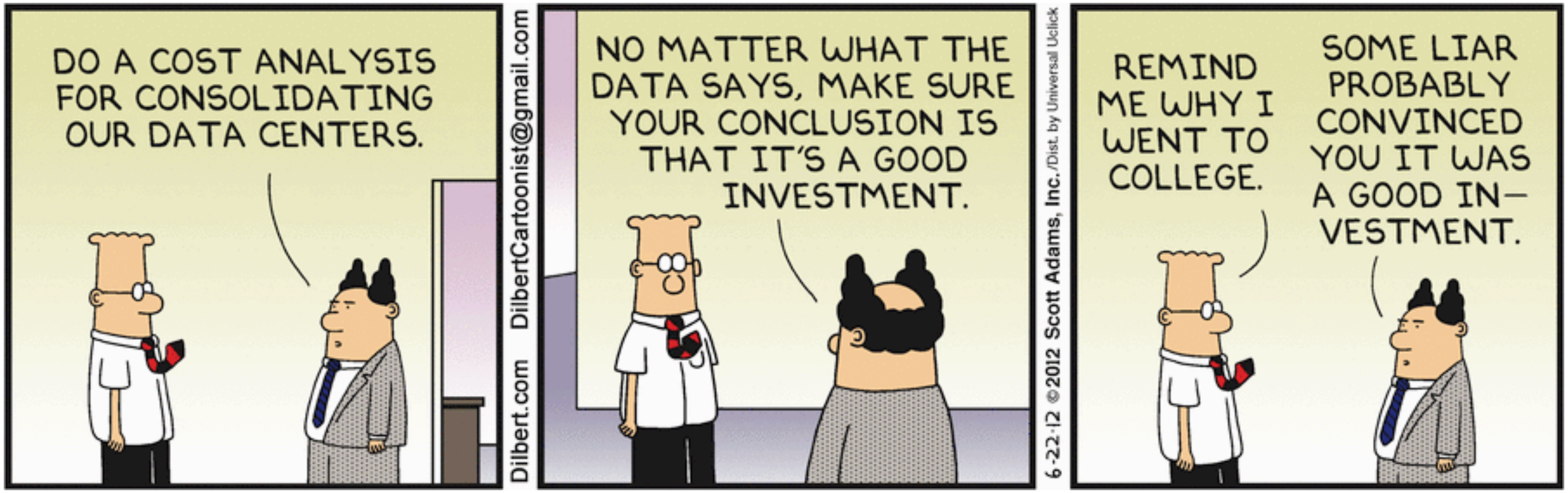
4. Build a model



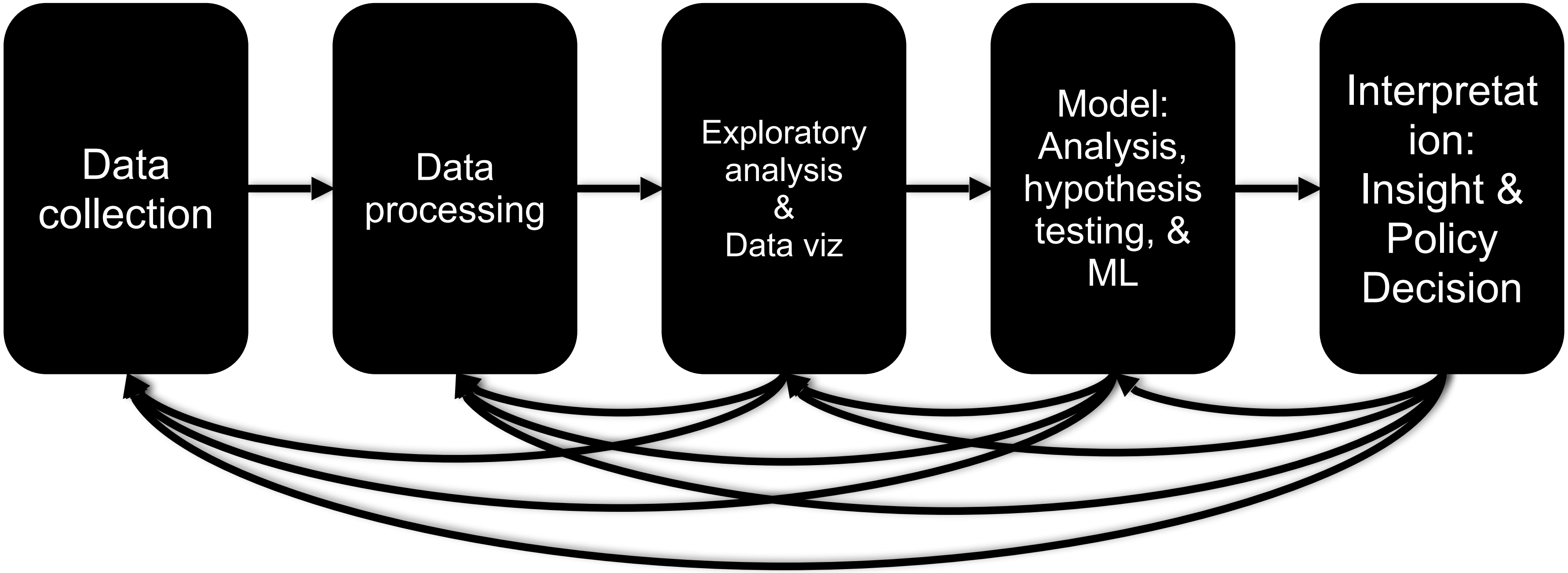
"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

DATA SCIENCE LIFECYCLE

5. Interpretation


DO A COST ANALYSIS FOR CONSOLIDATING OUR DATA CENTERS.
Dilbert.com DilbertCartoonist@gmail.com
NO MATTER WHAT THE DATA SAYS, MAKE SURE YOUR CONCLUSION IS THAT IT'S A GOOD INVESTMENT.
6-22-12 ©2012 Scott Adams, Inc./Dist. by Universal Uclick
REMIND ME WHY I WENT TO COLLEGE.
SOME LIAR PROBABLY CONVINCED YOU IT WAS A GOOD IN-VESTMENT.

THE DATA SCIENCE LIFECYCLE



Data collection

Data processing

Exploratory analysis & Data viz

Model: Analysis, hypothesis testing, & ML

Interpretation: Insight & Policy Decision

“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, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids.”

Hal Varian
Chief Economist at Google

MOTIVATION

Explosion of data, in pretty much every domain

- Sensing devices and sensor networks that can monitor everything 24/7 from temperature to pollution to vital signs
- Increasingly sophisticated smart phones
- Internet, social networks makes it easy to publish data
- Scientific experiments and simulations → astronomical data volumes
- Internet of Things
- Dataification: taking all aspects of life and turning them into data (e.g., what you like/enjoy has been turned into a stream of your "likes")

How to handle that data? How to extract interesting actionable insights and scientific knowledge?

Data volumes expected to get much worse

FOUR V'S OF BIG DATA

Increasing data Volumes

- Scientific data: 1.5GB per genome -- can be sequenced in .5 hrs
- 500M tweets per day
- 2.5 Quintillion bytes of data created every day

Variety:

- Structured data, spreadsheets, photos, videos, natural text, ...

Velocity

- Sensors everywhere -- can generate high-rate "data streams"
- Real-time analytics requires data to be consumed as fast as it is generated

Veracity

- How do you decide what to trust? How to remove noise? How to fill in missing values?

THIS COURSE

End-to-end data science lifecycle

Acquiring, wrangling, cleaning, and integrating data; Setting up pipelines for ETL

Data modeling

Information Visualization

Ethics, Privacy, and Reproducibility

Feel free to tell me if there are topics that you think we should cover…

Info: http://www.cs.umd.edu/class/fall2020/cmsc320-0201/
Piazza: https://piazza.com/class/ke8rnh7rg8t683
Gradescope: https://www.gradescope.com/courses/171922
ELMS: (everyone should be registered automatically)

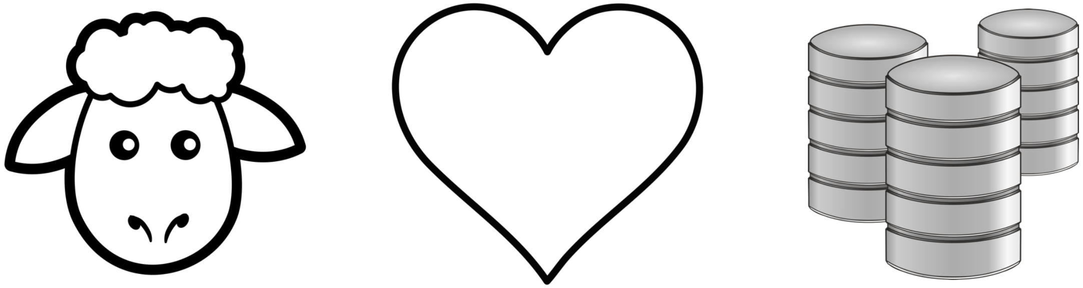
PREREQUISITE KNOWLEDGE

Aimed at folks with some CS knowledge – but likely accessible to others with programming experience and mathematical maturity.

We do not assume:

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

We do assume: You want to be here!



(TENTATIVE) COURSE STRUCTURE

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

Next : data collection & management

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

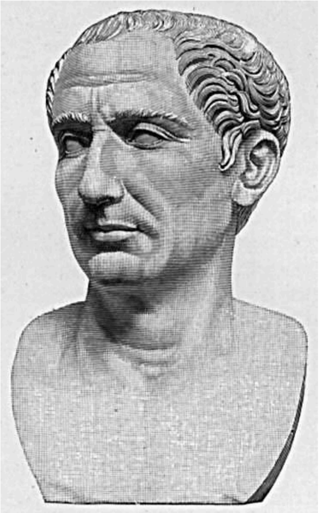
Next lectures: statistical modeling & ML

- Statistical learning, regression, classification, cross-validation, model evaluation, hypothesis testing, etc …

Midterm

Final lectures: advanced topics

- Dimensionality reduction, distributed learning, big data, distributed computation
- *M*ore lectures



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 and a pdf of that notebook on Gradescope.

GRADE #2: READING HOMEWORKS

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

- Blog posts
- Academic articles
- News articles

Weekly quiz to be taken in class, every Tuesday.

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: MIDTERM

You know what this is.

Will cover roughly the material up to the week before of class:

- Qualitative, and
- Quantitative

Currently scheduled for November 19th (let me know if you have hard constraints, etc)

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

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

FINAL TUTORIAL RUBRIC

It will be graded 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: John Dickerson

GRADE BREAKDOWN

40% mini-projects:

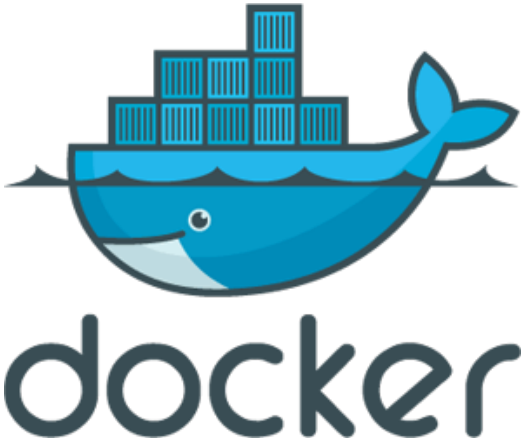
- There are 4 of them
- Equal weighting @ 10% each

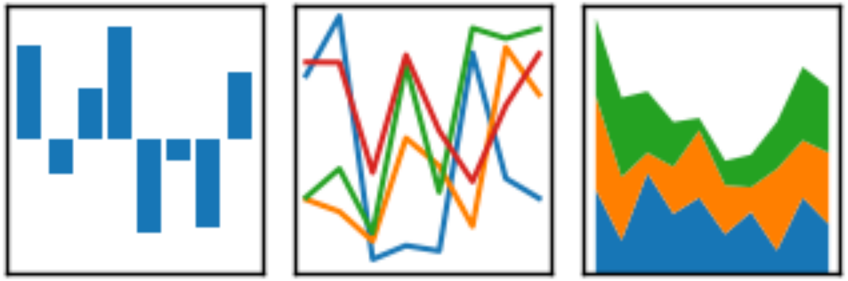
10% reading homeworks (Quizzes)

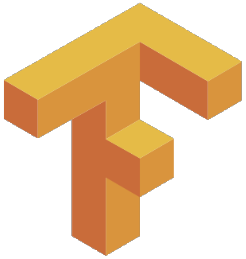
25% Midterm

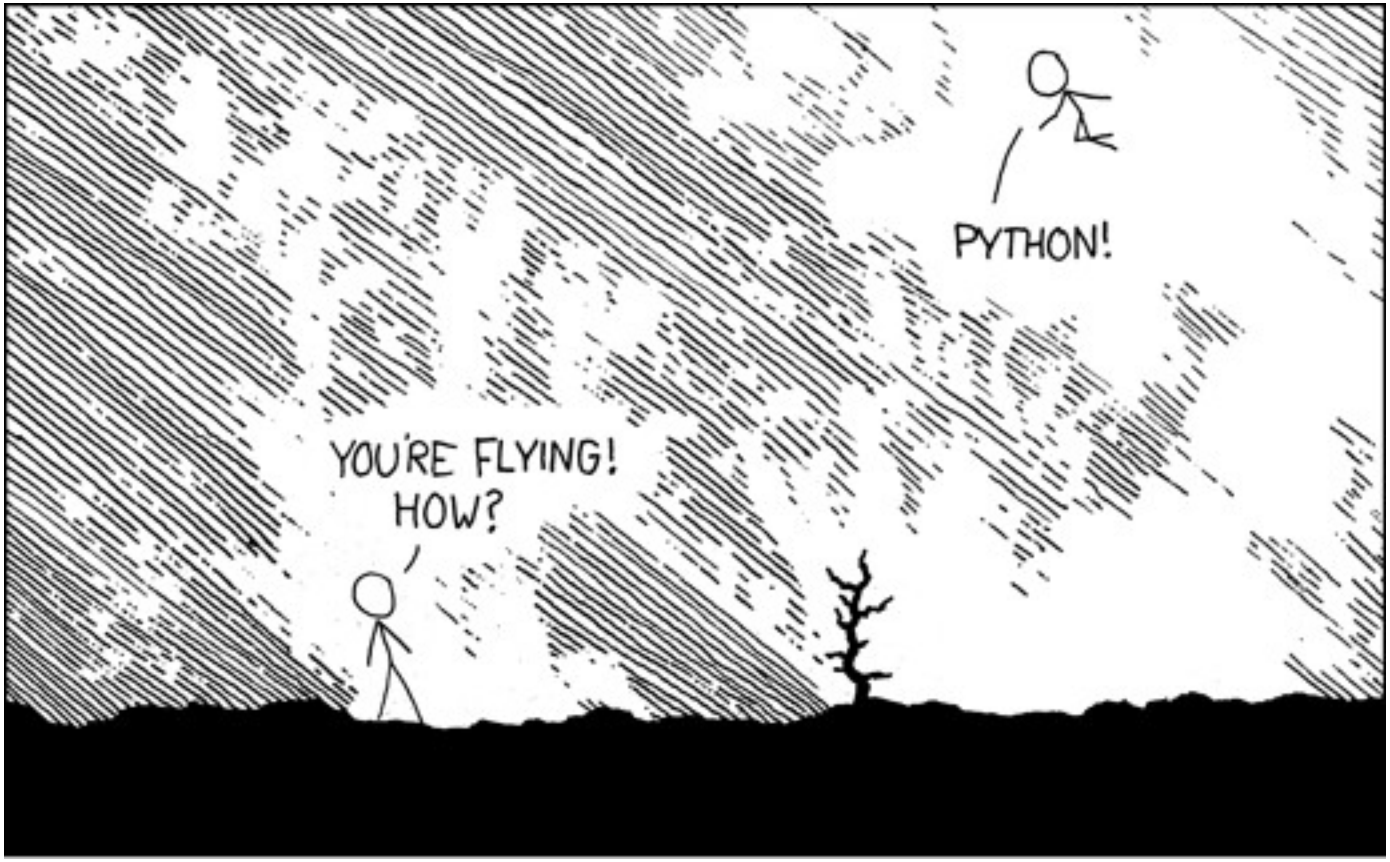
25% final tutorial

SOME TECHNOLOGIES WE WILL USE









(Don’t tell CMSC330 …)

IMPORTANT WALLS OF TEXT

Common Sense!

ANTI-HARASSMENT

(Adapted from ACM SIGCOMM's policies)

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.

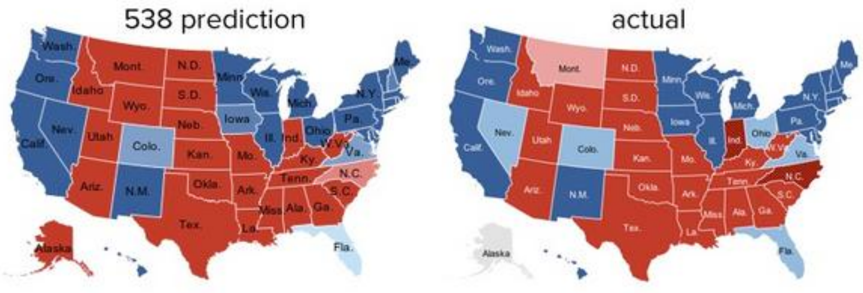
Everything you hand in must be in your own words, and based on your own understanding of the solution. If someone helps you understand the problem during a high-level discussion, you are not cheating. We strongly encourage students to help one another understand the material presented in class, in the 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

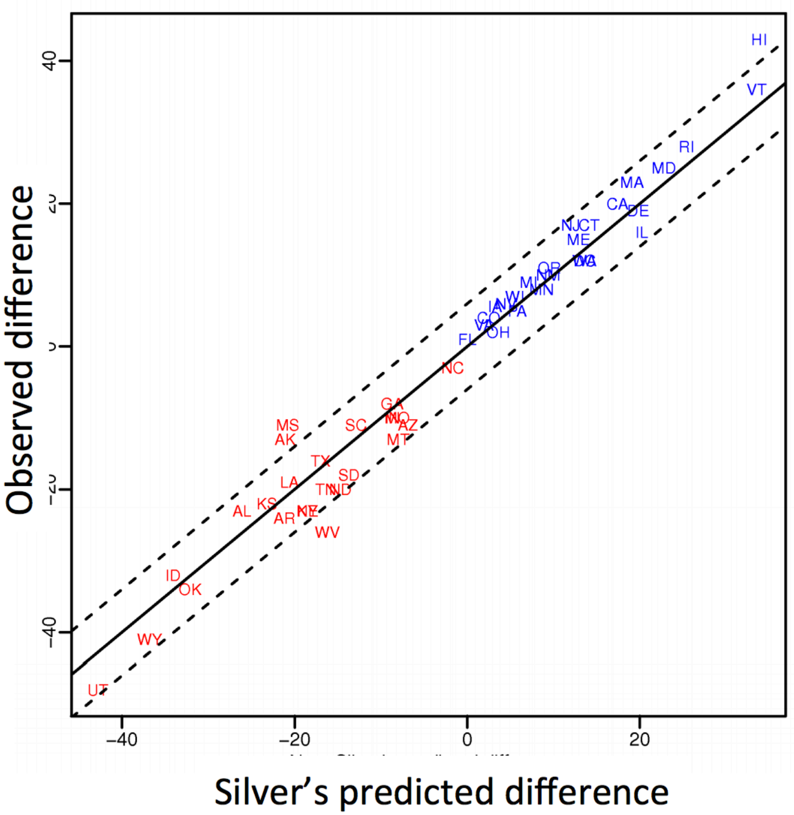
Nate Silver uses a simple idea – taking a principled approach to aggregating polling instead of relying on punditry – and:

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



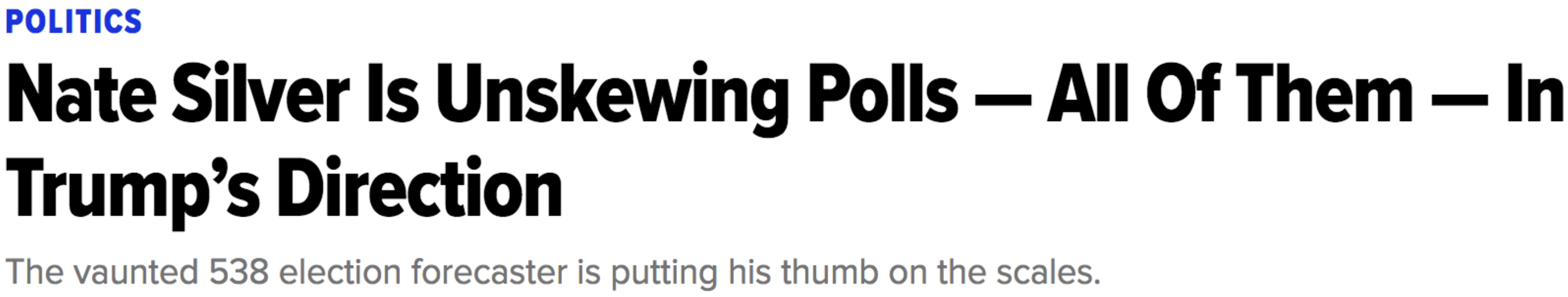
- **(He is also a great case study in creating a brand.)**

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



Democrat (+) or Republican (-) in 2012

POLLING: 2016



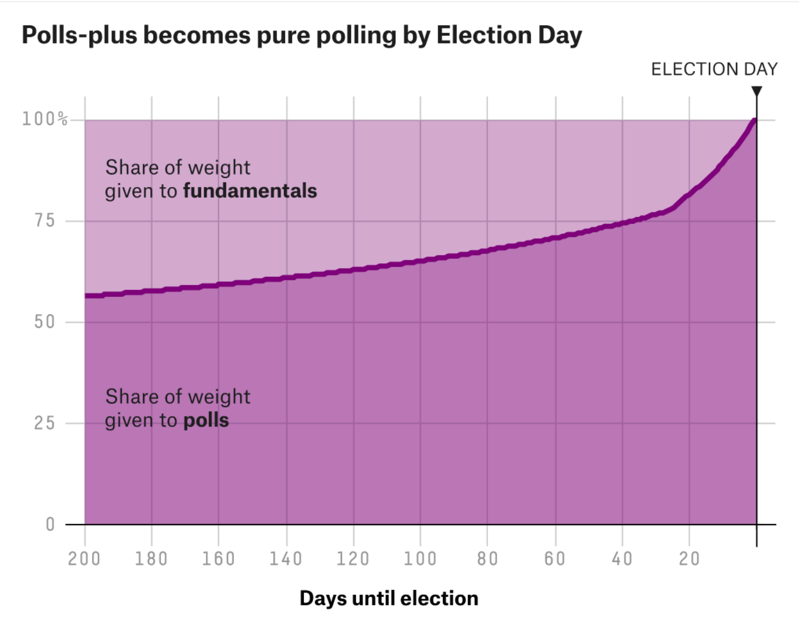
POLITICS

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



Polls-plus becomes pure polling by Election Day

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

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, Dressel and Farid 2018]

textio

DOXA

gild

GJ GapJumpers

zest finance

LendingClub

Kabbage

PREDPOL

NORTHPOINTE
Advancing Justice. Embracing Community.

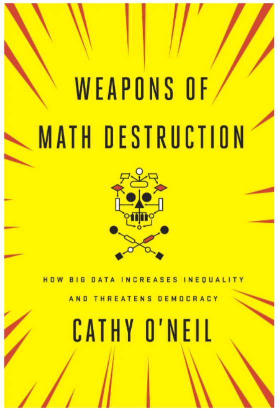
Hiring

Lending

Policing/ sentencing

Search for minority names →
ads for DUI/arrest records

Female cookies →
less freq. shown professional job opening ads



WEAPONS OF MATH DESTRUCTION

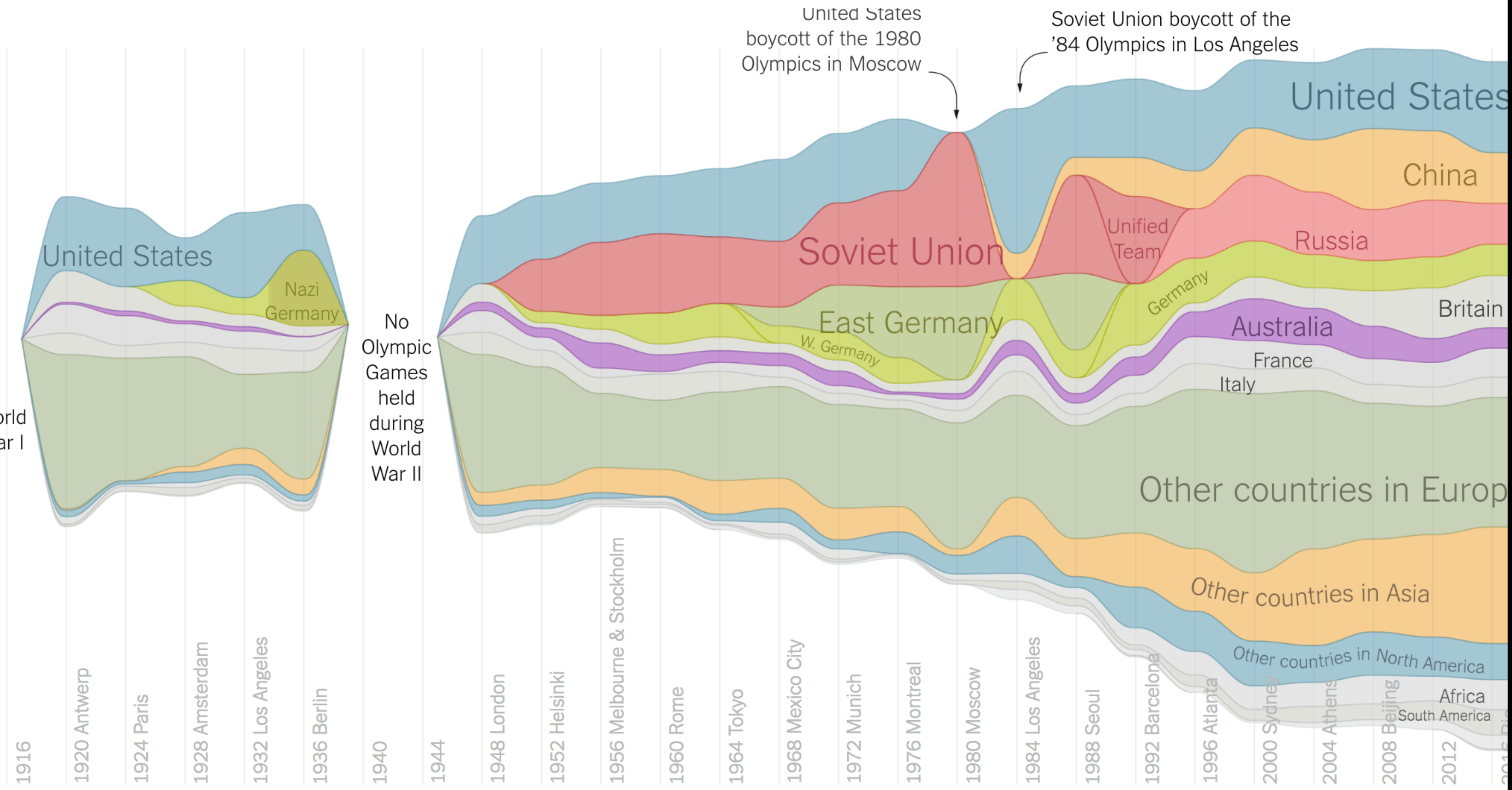
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

"… 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]

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%

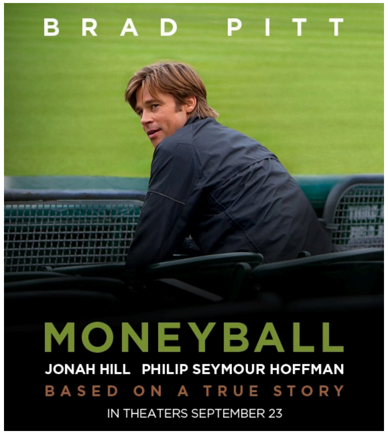
- **Happened after about three years**
- **Model was never used by Netflix for a variety of reasons**
  - Out of date (DVDs vs streaming)
  - Too complicated / not interpretable

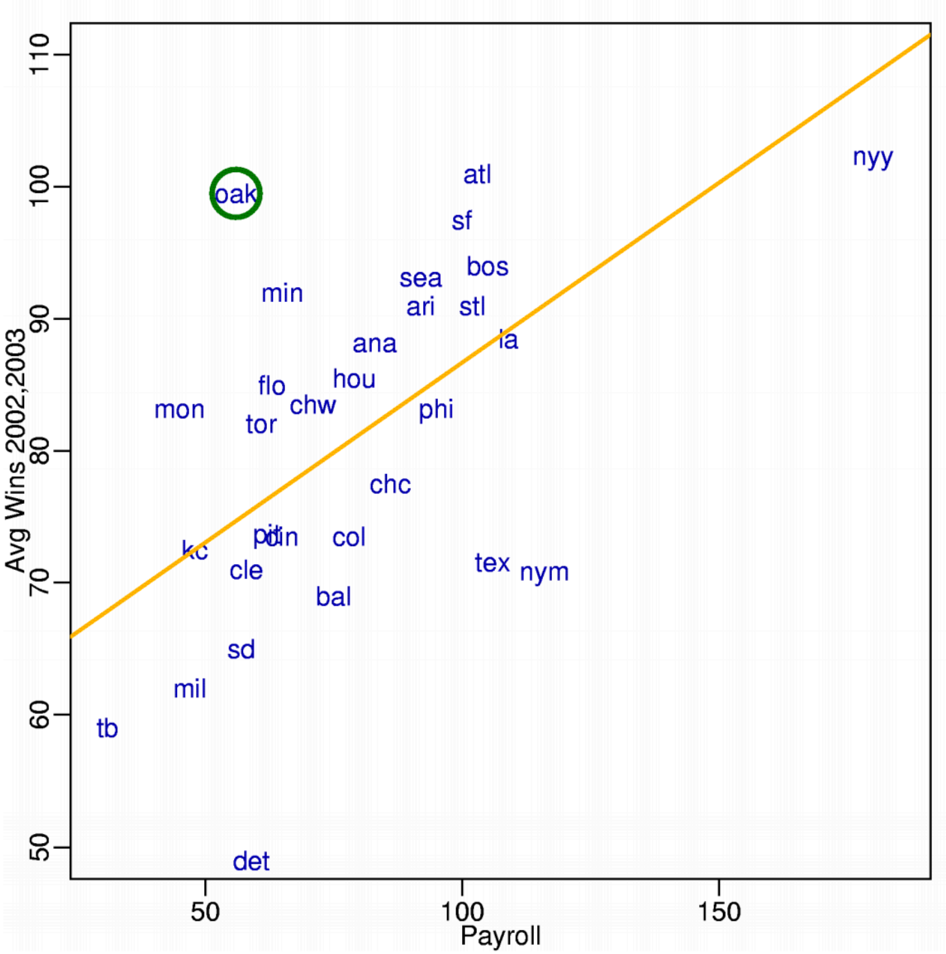
MONEYBALL

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

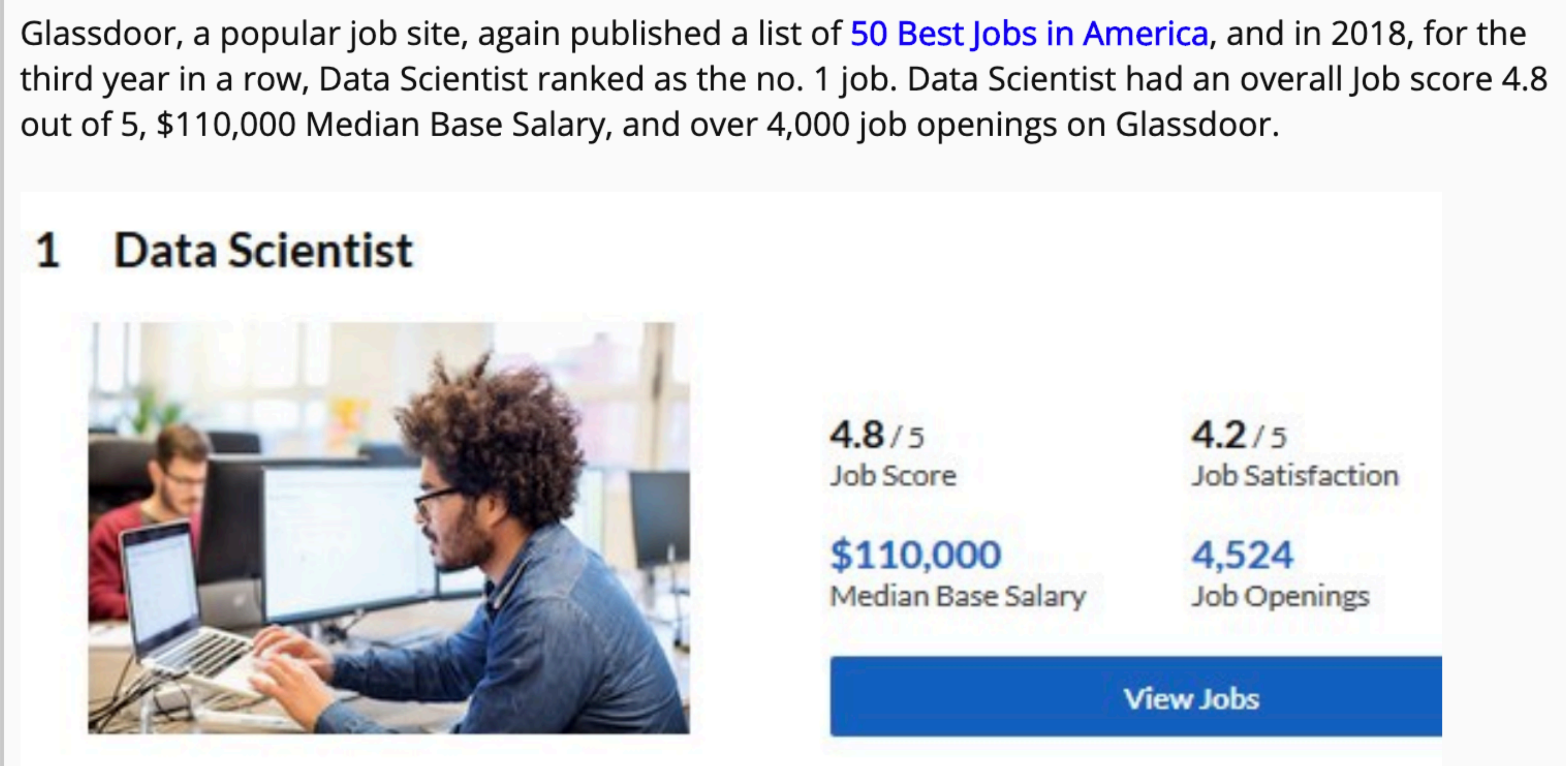
Billy Bean, data scientist *du jour*, convinces the {bad, poor} Oakland Athletics to use a quantitative aka sabermetric approach to hiring

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





DATA SCIENTIST



Glassdoor, a popular job site, again published a list of 50 Best Jobs in America, and in 2018, for the third year in a row, Data Scientist ranked as the no. 1 job. Data Scientist had an overall Job score 4.8 out of 5, $110,000 Median Base Salary, and over 4,000 job openings on Glassdoor.

1 Data Scientist

4.8 / 5
Job Score

4.2 / 5
Job Satisfaction

$110,000
Median Base Salary

4,524
Job Openings

WRAP-UP FOR PART I

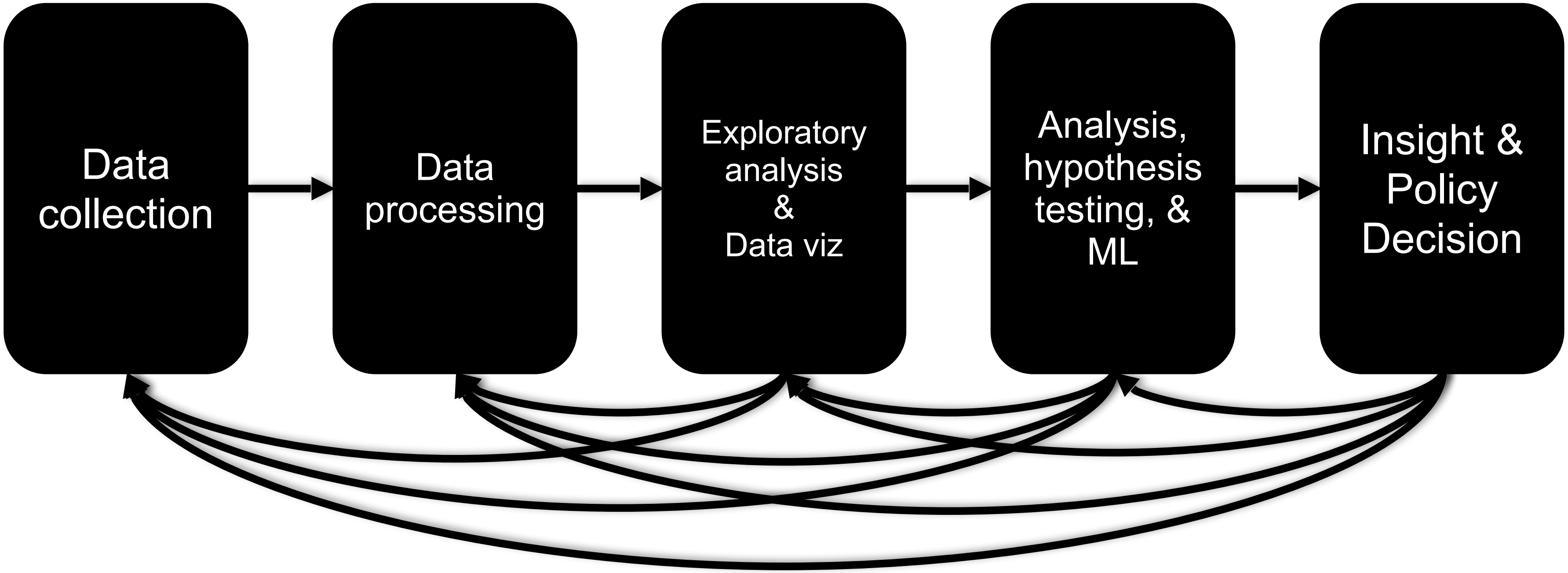
Please chat with me if you’re unsure of whether or not you’re at the right {programming, math} level for this course:

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

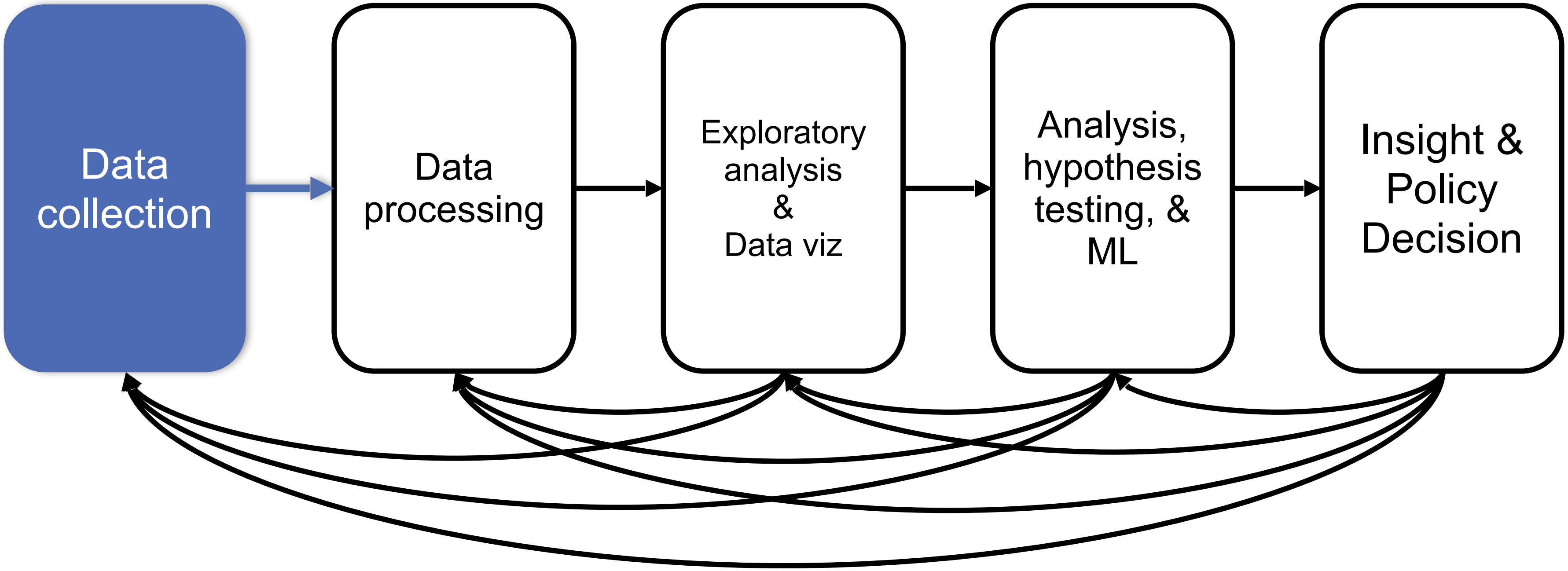
Read about Docker & Jupyter!

- Works on *nix, OSX, Windows
- https://www.docker.com/
- (Project 0 is posted.)

THE DATA LIFECYCLE


Data collection
Data processing
Exploratory analysis & Data viz
Analysis, hypothesis testing, & ML
Insight & Policy Decision

TODAY'S LECTURE


Data collection
Data processing
Exploratory analysis & Data viz
Analysis, hypothesis testing, & ML
Insight & Policy Decision

BUT FIRST, SNAKES!

python™

Python is an interpreted, dynamically-typed, high-level, garbage-collected, object-oriented-functional-imperative, and widely used scripting language.

- **Interpreted:** instructions executed without being compiled into (virtual) machine instructions*
- **Dynamically-typed:** verifies type safety at runtime
- **High-level:** abstracted away from the raw metal and kernel
- **Garbage-collected:** memory management is automated
- **OOFI:** you can do bits of OO, F, and I programming

Not the point of this class!

- Python is fast (developer time), intuitive, and used in industry!

*you can compile Python source, but it's not required

THE ZEN OF PYTHON

- Beautiful is better than ugly.
- Explicit is better than implicit.
- Simple is better than complex.
- Complex is better than complicated.
- Flat is better than nested.
- Sparse is better than dense.
- Readability counts.
- Special cases aren't special enough to break the rules …
- … although practicality beats purity.
- Errors should never pass silently …
- … unless explicitly silenced.

Thanks: SDSMT ACM/LUG

LITERATE PROGRAMMING

Literate code contains in one document:

- the source code;
- text explanation of the code; and
- the end result of running the code.

Basic idea: present code in the order that logic and flow of human thoughts demand, not the machine-needed ordering

- Necessary for data science!
- Many choices made need textual explanation, ditto results.

Stuff you'll be using in Project 0 (and beyond)!

IP[y]: IPython Interactive Computing

jupyter