Introduction to Data Science

Nicholas Mattei, *Tulane University*

CMPS3660 – Introduction to Data Science – Fall 2019

https://rebrand.ly/TUDataScience

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

Data science comprises three distinct and overlapping areas: the skills of a *statistician* who knows how to model and summarize datasets; the skills of a *computer scientists* who can design and use algorithms to efficiently store, process, and visualize the data; and the *domain expertise* necessary to formulate the right question and put the answers in context

Jake VanderPlas
Author, *Python Data Science Handbook*

The goal of data science is to improve decision making based on insights from data. As a field, Data Science encompasses a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from datasets.

John D. Kelleher and Brendan Tierney
Authors, *Data Science (MIT Press)*

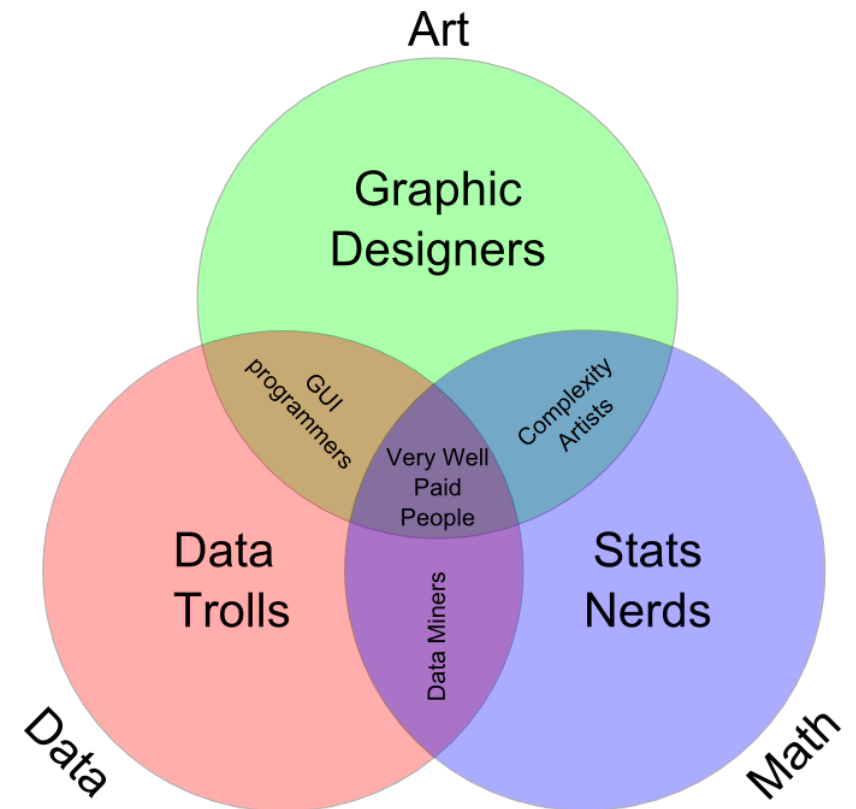- The battle of the Data Science Venn Diagrams -- https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html
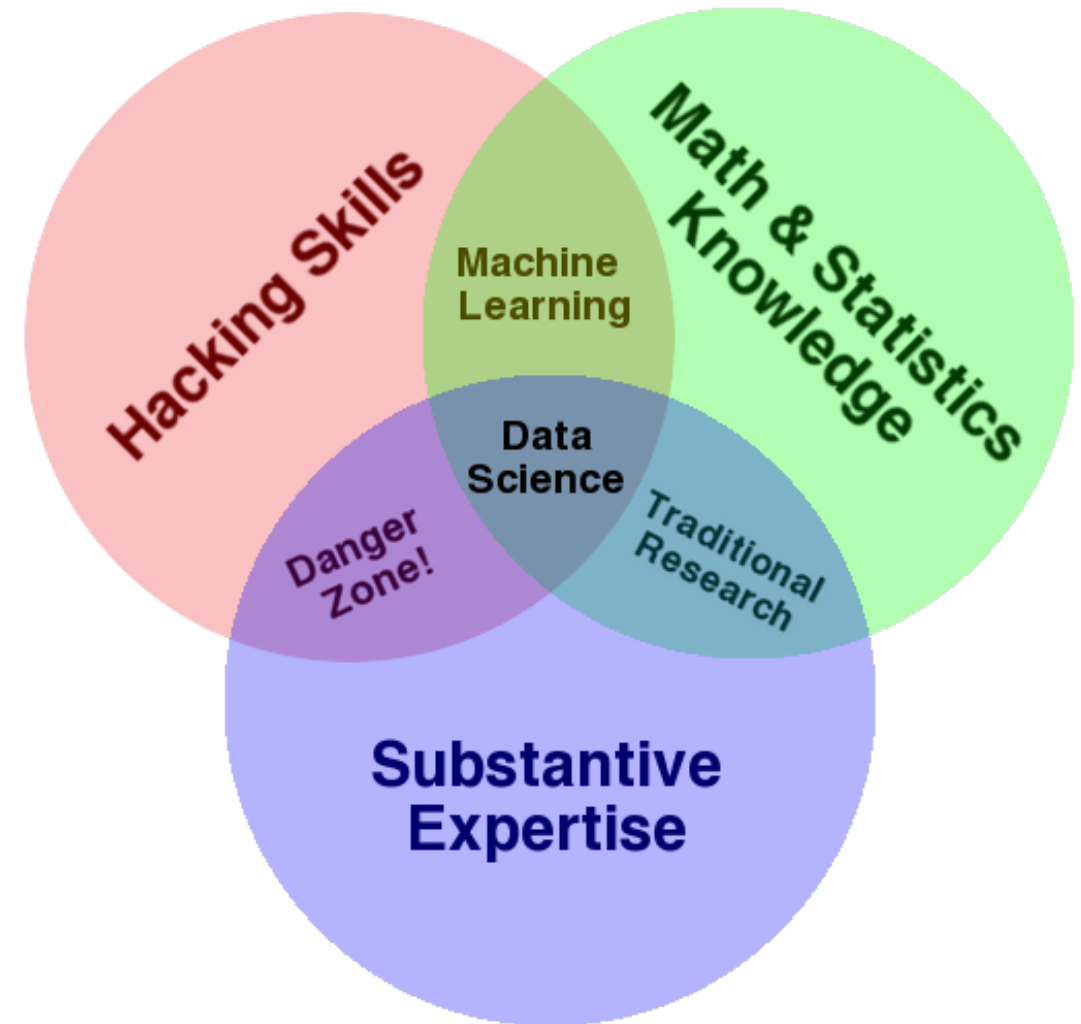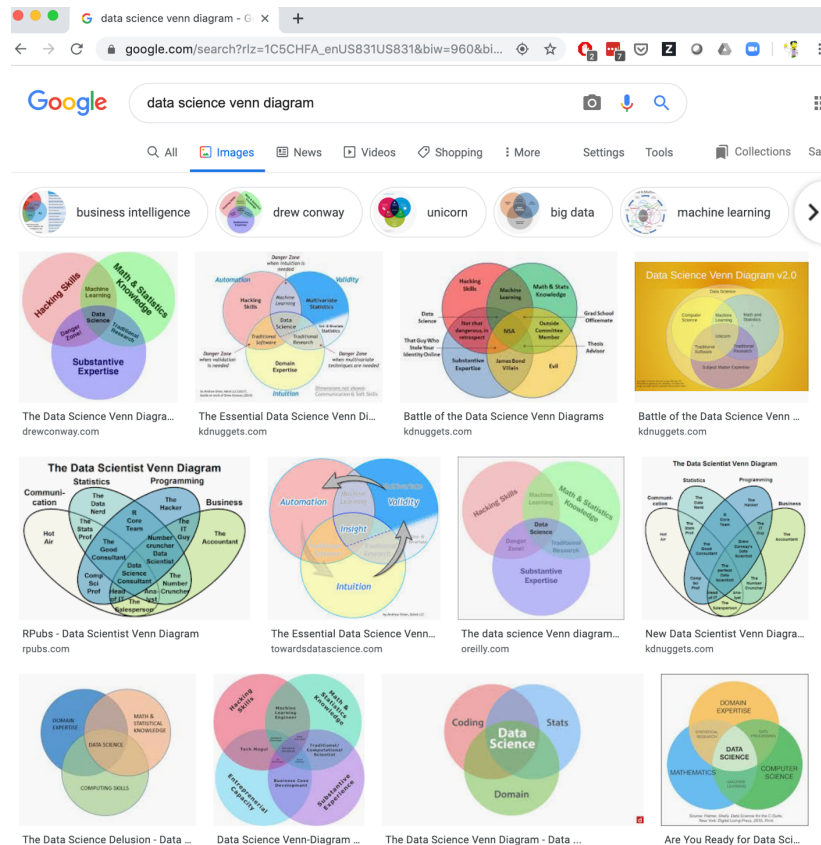
Drew Conway
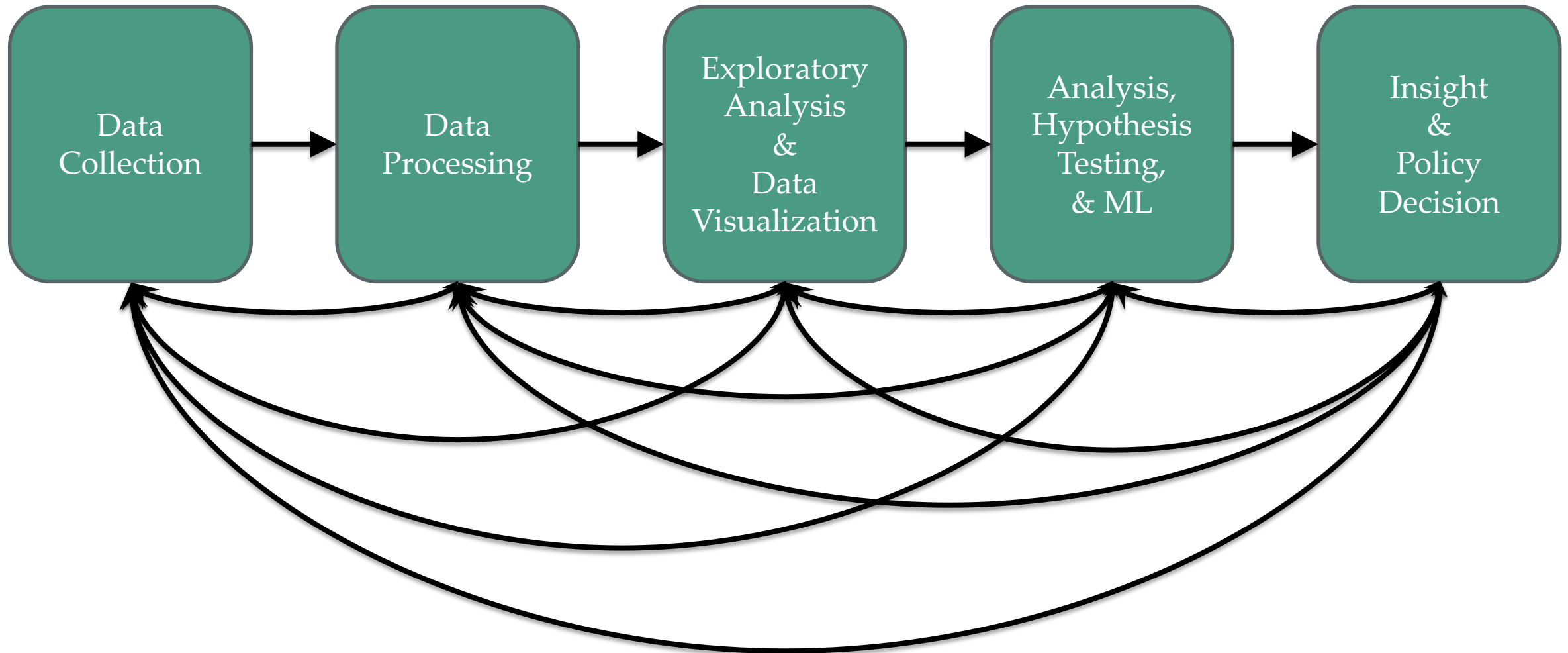CEO, Alluvium (analytics company)

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

Art

Graphic Designers

GUI programmers

Complexity Artists

Very Well Paid People

Data Trolls

Data Miners
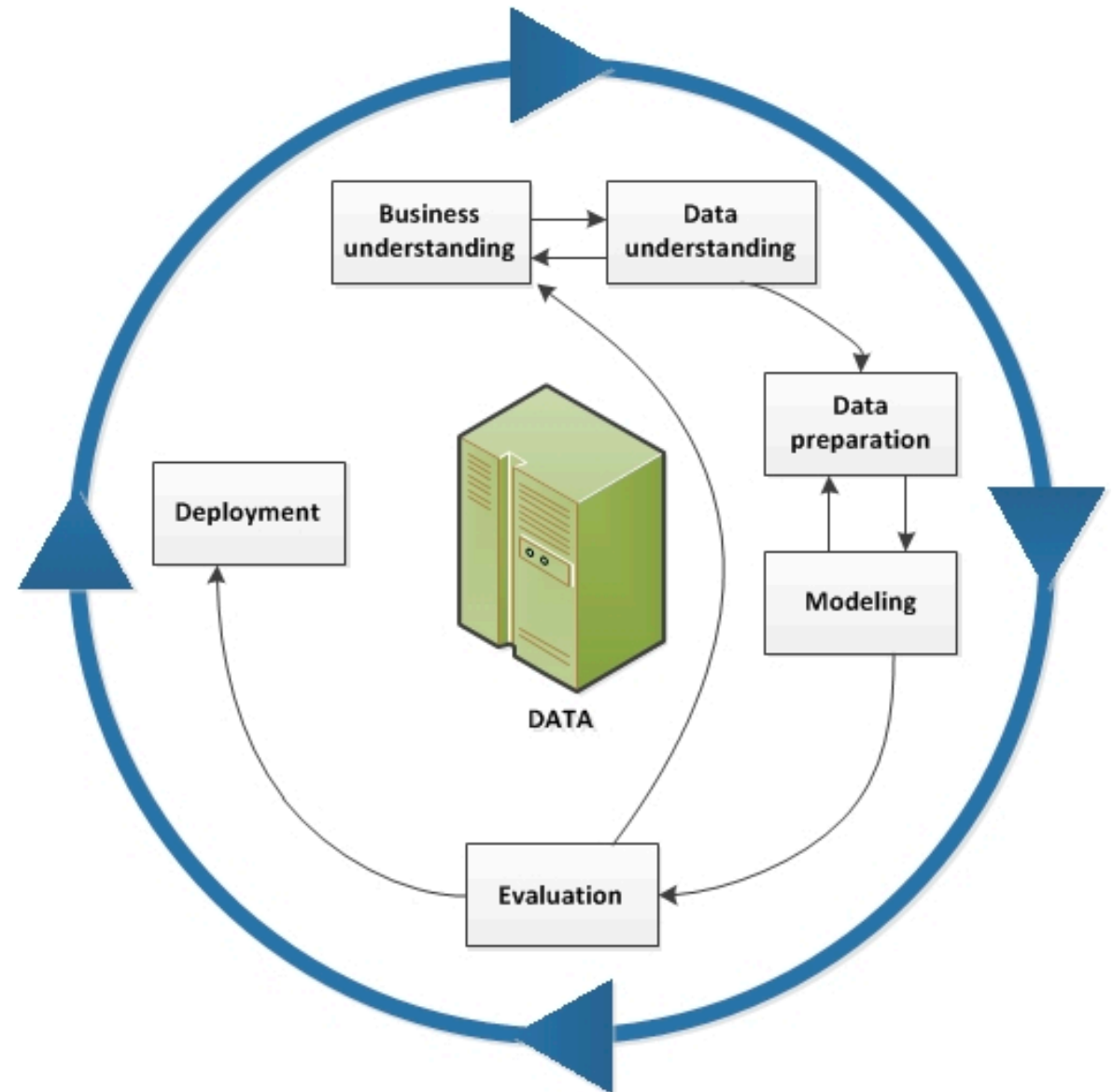
Stats Nerds

Data

Math

Tuomas Carsey, UNC

The Data LifeCycle

The CRISP-DM LifeCycle

- From IBM Data LifeCycle:
  https://www.ibm.com/support/knowledgecenter/SS3RA7_18.2.0/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview_container.html

- Going around around (sometimes these projects never end, e.g., Tugboats.)…

Data Science Pyramid

- Adapted from *Data Science,* Kelleher and Tierney – Taken from *Data Mining* by Han, Kamber, and Pei.

- Related to the DIKW Pyramid - https://en.wikipedia.org/wiki/DIKW_pyramid
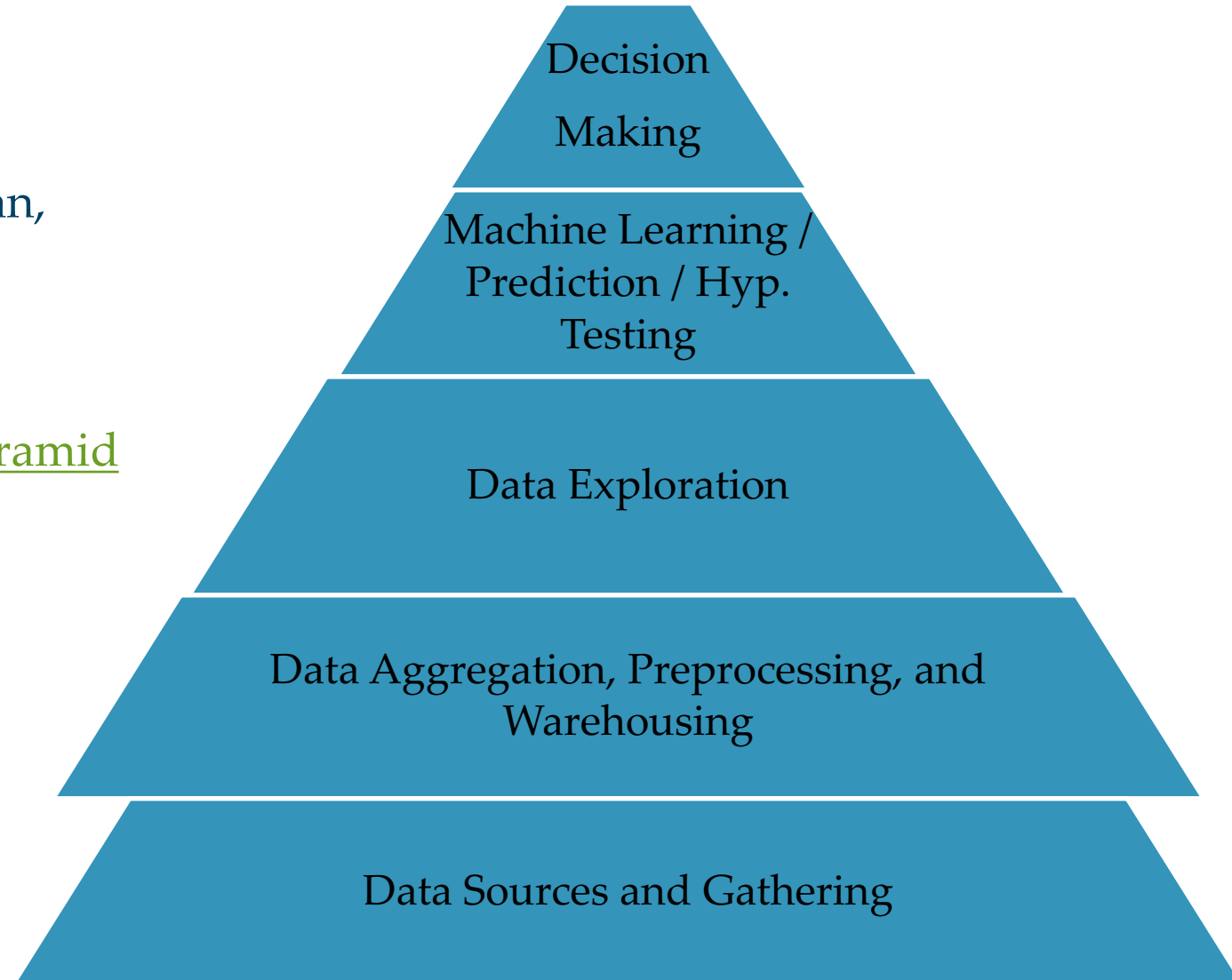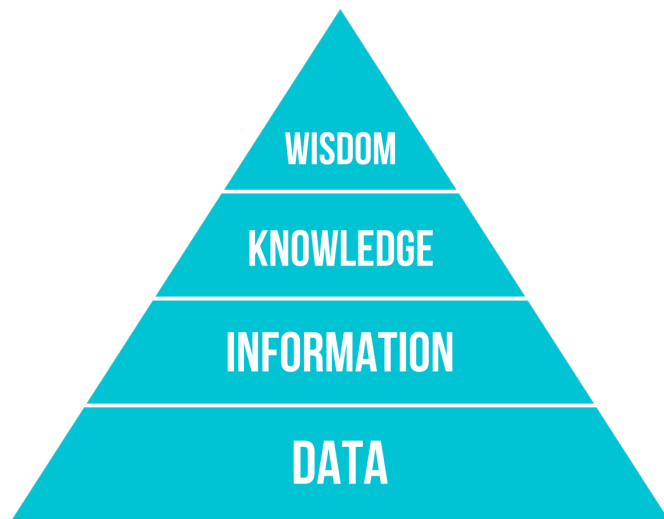
"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids."

Hal Varian
Chief Economist at Google

Hal Varian

This Course

- You'll learn to take data:
    - Process it; Visualize it; Understand it; Communicate it; Extract value from it
- At the conclusion of this course students will be able to:
    - Open, load, and manipulate data using industry standard tools.
    - Have a basic understanding of data management and storage systems.
    - Be able to clean data and perform basic statistical analysis of the data including visualization.
    - Have an understanding of statistical hypothesis testing including t-tests and bootstrapped confidence intervals.
    - Be able to use one or more machine learning algorithms for classification, regression, and prediction of data.
    - Be able to present the results of a complete data analysis in written, visual, and presentation.

Overview

- Course Webpage:
  - **https://rebrand.ly/TUDataScience**
  - We will use this post all course material including projects, readings, and slides.
  - Quizzes and projects will be turned in through Canvas / GitHub.

- Time and Place
  - **Lectures:** Tuesdays and Thursdays
  - **Room:** Stanly Thomas 302 (Building 10)
  - **Time:** 1100 – 1215

Course Staff

- **Instructor:** Prof. Nicholas Mattei

  – **Office:** Stanley Thomas 402B

  – **Office Hours:** T 1400-1500 and TR 1600-1700

  – Email: nsmattei@tulane.edu

  – **Research Interests:** AI, ML, data science, decision making… (this class!)

- **TA:** Arie Glazer

  – Email: aglazier@tulane.edu

  – Office: Stanley Thomas 309
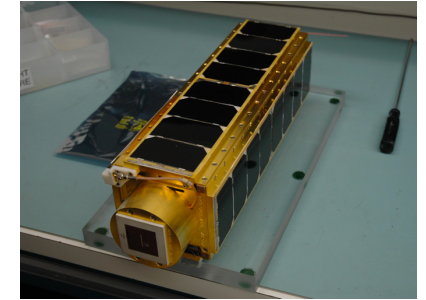
  – Office Hours: MW 1400-1500

Who am I?

- Undergrad/Masters University of Kentucky

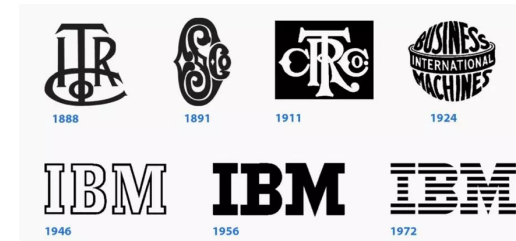- Worked at NASA Ames Research Center (Bay Area)

- PhD University of Kentucky

- NICTA / UNSW / Data61 / CSIRO (Sydney, Australia)

- IBM Research (New York)

- Tulane University!

# This Course is an EXPERIMENT!

- Please speak up if things aren't working, we're going to fast, or we're going too slow.
  - Trying some new things like in class Lab sessions!

- I have a tendency to speak fast, *please tell me to slow down!*

- The schedule is subject to change – please keep checking and I'll remind you of any changes!

- **Most important:** *let's have fun and learn some cool stuff!*

Prerequisite Knowledge

- Aimed at **TU:CS Undergrads** – but likely accessible to others with programming experience and mathematical maturity.

- We **do not** assume:
  - Extensive experience with Python, pandas, scikit-learn, matplotlib, etc …
  - Deep statistics or any ML knowledge
  - Database or distributed systems knowledge

- We **do** assume:
  - Some Familiarity with Python and/or programming.
  - You are comfortable with topics in Discrete Math (sets, probability, counting).
  - You want to be here!
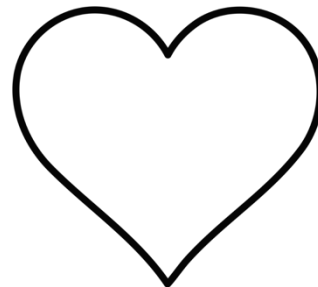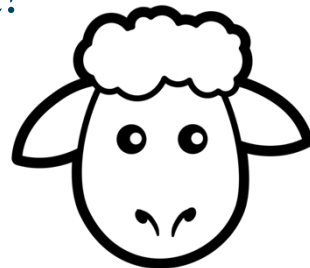
Some Technologies we will use
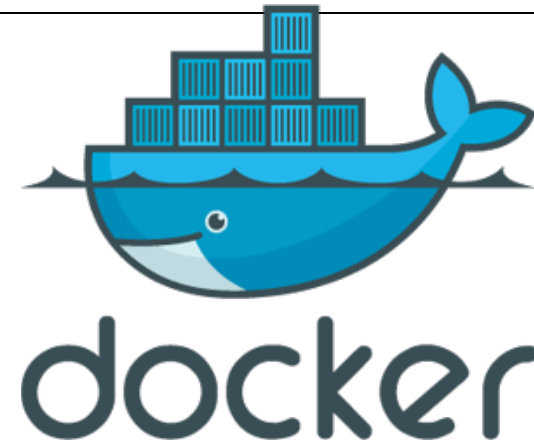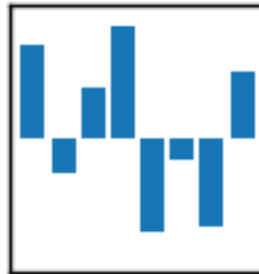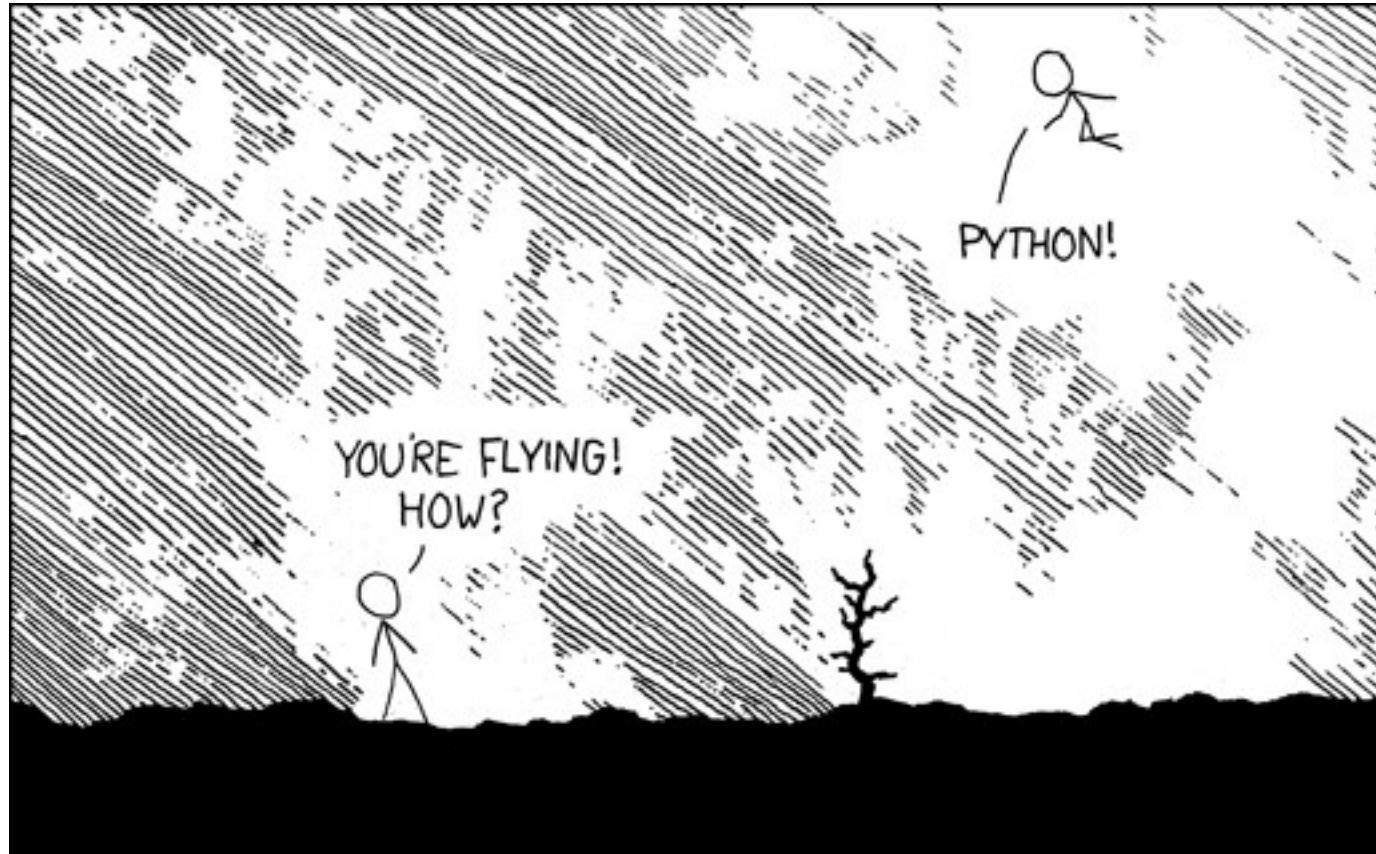
You really don't need anything else…

Who are you? -- https://rebrand.ly/DSQuestions

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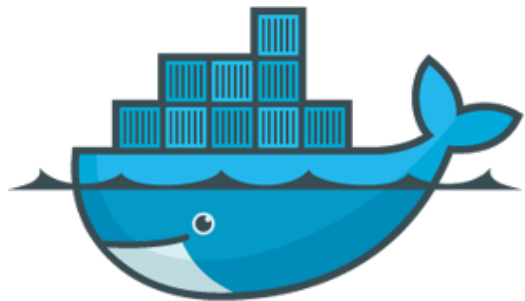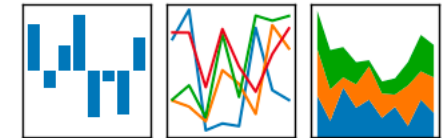
Required Materials

- There is no required textbook for this course. However, we will make extensive use of online textbooks and articles for the required reading that you will be quizzed on.
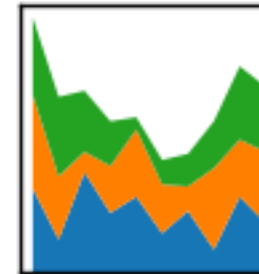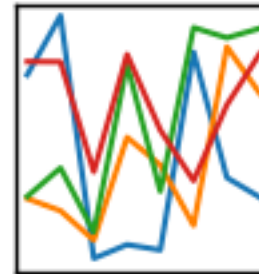- You will also need access to a computer that you can install the required software (Anaconda, Docker). **If you do not have access to a computer that you can install software on please see me ASAP.**

- **Online Books:**
  - *Python Data Science Handbook: Essential Tools for Working with Data,* Jake VanderPlass. O'Reilly Media Inc., 2016. Available online for free at: https://github.com/jakevdp/PythonDataScienceHandbook
  - *Computational and Inferential Thinking: The Foundations of Data Science,* Ani Adhikari and John DeNero. A free online textbook that includes interactive Jupyter notebooks and public data sets for all examples at: https://www.inferentialthinking.com/chapters/intro
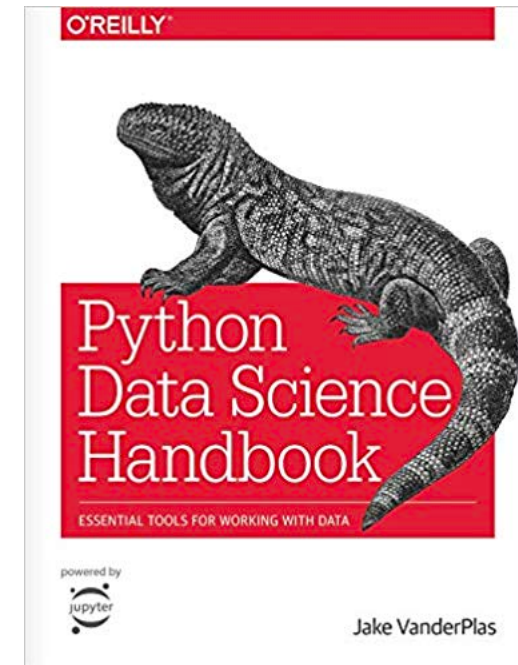
(Tentative) Course structure

- First Lecture: Intro!

- Next 4 lectures (2 weeks): intro & primers in the Python data science stack and tools

- Next 7 lectures (3.5 weeks): data collection & management
  - Best practices, data wrangling, exploratory analysis, ethics, debugging, visualization, etc …

- **Midterm**

- 4 Lectures (2 weeks) after fall break: basic probability, statistical testing, bootstrap.

- 9 Lectures (4.5 weeks): Machine learning tools, NLP, regression, classification etc.

- 2 Lectures (1 week): In class presentations of projects.
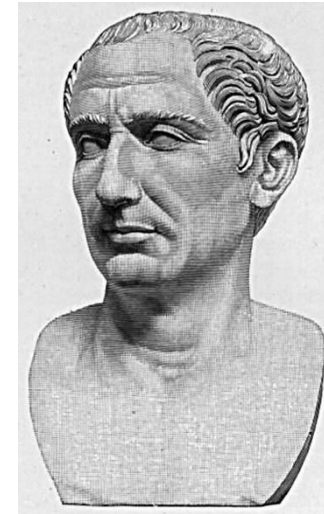
Ambitious …

- We have 29 total lectures.
  - Dr. Mattei gone for 3 (1 Midterm)
  - Class Presentations for 2
- We have 24 (-1 for intro) content lectures and so many things we can cover!!

Grading

- 15% - (Mostly) Weekly Questions
- 40% - Mini-Projects and Labs
- 15% - Midterm Exam
- 30% - Final Project

- All grades will be posted on Canvas.

A >= 90%   B >= 80%   C >= 70%

D >= 60%   F < 60%

+/- grades will be given for borderline cases.

Grade Breakdown

Questions   Mini-Projects   Midterm   Final Project
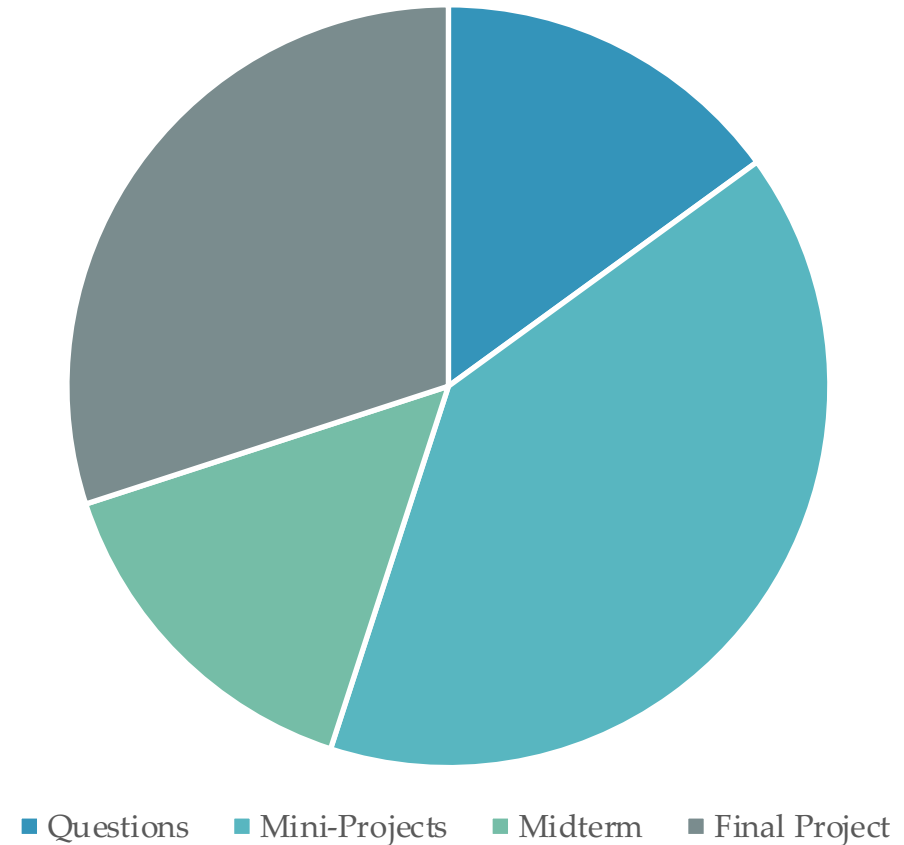
15% - Weekly Questions

- ~10 - 12 Weekly Questions

  – Will be assigned on Canvas.

  – Will cover posted articles, tutorials, and in class lectures.

  – Will be short answer and multiple guess.

  – Typically will be assigned on Tuesdays and due by Midnight the following Tuesday.

  – No late work allowed.

- Requests for an extension (with a valid reason) must be given to the instructor before the questions are due.

Mini-Projects and Labs (40%)

- Students will complete mini-project and lab assignments:
  - **Exercises** showing you are comfortable with the data science stack.
    - Typically in class Labs!
  - **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
  - **Also Never Allowed:** paying someone to do your work!

- Deliverable:
  - Turn in an .ipynb of a Jupyter notebook on Canvas
  - First project is already posted (getting Github and your environment setup).

Midterm Exam (15%)

- You know what this is.

- Will cover roughly the first 3/5 of class:
  - Qualitative (more)
  - Quantitative (less)

- You are allowed one, hand written, 8.5x11in, cheat sheet. This will be turned in with your exam (and handed back).

Final Tutorial (30%)

- In lieu of a final exam, you'll (individual or group) 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
- Deliverables (we will have several milestones as the semester goes):
  - URL of your own GitHub Pages site hosting an .ipynb/.html export of your final tutorial.
  - 10 min in class presentation of the tutorial (TBD based on number of groups).
- 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
- Example writeup:
  - https://andresgogo.github.io/

Final Tutorial Milestones (Running Notebook)

1. Partners and Datasets (Oct. 1)

   – 1 page writeup listing a partner (if any) and one to three datasets that you are considering working with and why. This is just an outline to make sure you are thinking. This will be published on your GitHub IO page so this also makes sure you've figured out how to get it uploaded!

2. Extraction, Transform, and Load (ETL) + Exploratory Data Analysis (EDA) (Oct. 22).

   – Your notebook from Part 1 but expanded to include the data being loaded and showing that you have figured out how to get the data into your system.

   – Your notebook expanded to include some graphs, visualizations, and stats that show you can manipulate your data and understand the dataset you are working with.

3. Final Presentation (Dec 3 + Dec 5).

   – Your in class presentation.

4. Final Tutorial (???)

   – Your posted final tutorial.

Final Tutorial Rubric (More info to come...)

- 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?

  - **Presentation**: Is the in class presentation well constructed? did you identify all the important bits of your tutorial? was the presentation engaging?

Some Data Set Options…

- 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, NOAA, …)
- https://www.kaggle.com/datasets
    – 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, …)
- Data is Plural: https://tinyletter.com/data-is-plural
    – https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juclhjFgqIY8fQFMemwKL2c64vk/edit#gid=0 - Lots of interesting and strange datasets.
- Corgies Datasets: https://think.cs.vt.edu/corgis/
- GitHub Awesome Public Datasets: https://github.com/awesomedata/awesome-public-datasets
- Think Stats Datasets: https://docs.google.com/spreadsheets/d/e/2PACX-1vQlv2BpO7TsU2UAE7iQwCUxvxn9LTXEPPj5EnA0l9-DJSwCDZU0xg_IhWBItZ7bNvZ_BznznuOrYYy0i/pubhtml

Class Attendance and Participation

- Class attendance

  – Class attendance will be taken almost every day.

  – **Please please please do the readings (if possible) before so we can talk about them!**

  – Students are responsible for notifying instructors about absences that result from serious illnesses, injuries, or critical personal problems

  – Students with frequent absences will be reported and/or removed from the course according to university policy.

Electronic Devices

- Please silence your phones!

- If you want to use a laptop please move to the back of the class
  - https://www.scientificamerican.com/article/students-are-better-off-without-a-laptop-in-the-classroom/

- **<u>EXCEPTIONS!</u>**
  - We will have questions online like today at several points (live graphs!)
  - We will have some days where I encourage you to bring a laptop and follow along.

Giant and Important Walls of Text

Academy Integrity

- You are required to adhere to Tulane's Code of Academic Conduct.

- This course will follow Tulane's Code of Academic Conduct. Cheating will be reported to the Associate Dean of Newcomb-Tulane College. Discussion is encouraged. However, what you turn in must be your own. You may not read another classmate's solutions or copy a solution from the web. **I will be running checks on the code turned in for plagiarism. If plagiarism is detected the minimum penalty is a 0 on the assignment and being reported, however, you may automatically fail this course at my discretion.**

- The Code of Academic Conduct applies to all undergraduate students, full-time and part-time, at Tulane University. Tulane University expects and requires behavior compatible with its high standards of scholarship. By accepting admission to the university, a student accepts its regulations (i.e., Code of Academic Conduct and the Code of Student Conduct) and acknowledges the right of the university to take disciplinary action, including suspension or expulsion, for conduct judged unsatisfactory or disruptive.

Academic Integrity

- **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.
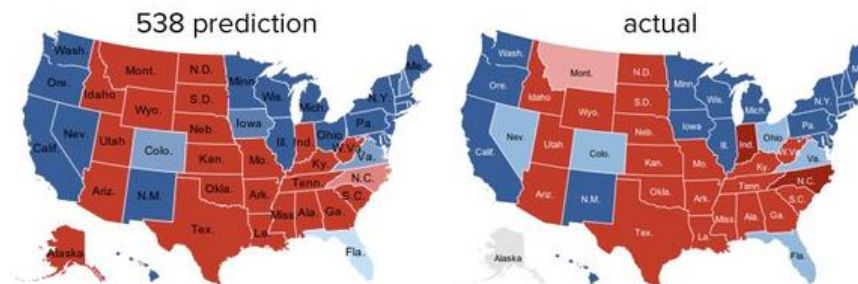
Title IX and Harassment

- Tulane University recognizes the inherent dignity of all individuals and promotes respect for all people. **As such, Tulane is committed to providing an environment free of all forms of discrimination including sexual and gender-based discrimination, harassment, and violence like sexual assault, intimate partner violence, and stalking. If you (or someone you know) has experienced or is experiencing these types of behaviors, know that you are not alone.**

- Resources and support are available: you can learn more at http://allin.tulane.edu. Any and all of your communications on these matters will be treated as either "Confidential" or "Private". **Please know that if you choose to confide in me I am mandated by the universISTY to report to the Title IX Coordinator, as Tulane and I want to be sure you are connected with all the support the university can offer.** You do not need to respond to outreach from the university if you do not want. You can also make a report yourself, including an anonymous report, through the form http://tulane.edu/concerns.

**(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
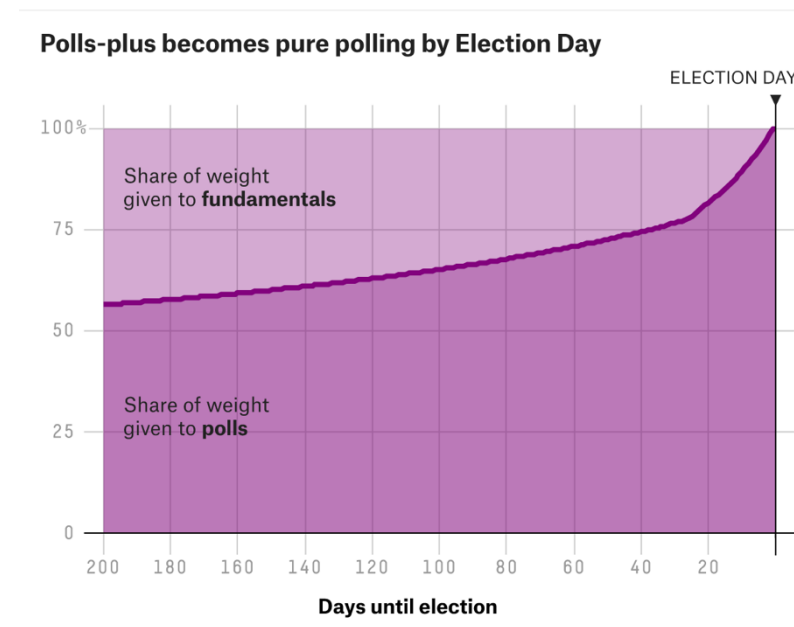

- (He is also a great case study in creating a brand.)

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

Democrat (+) or Republican (-) in 2012
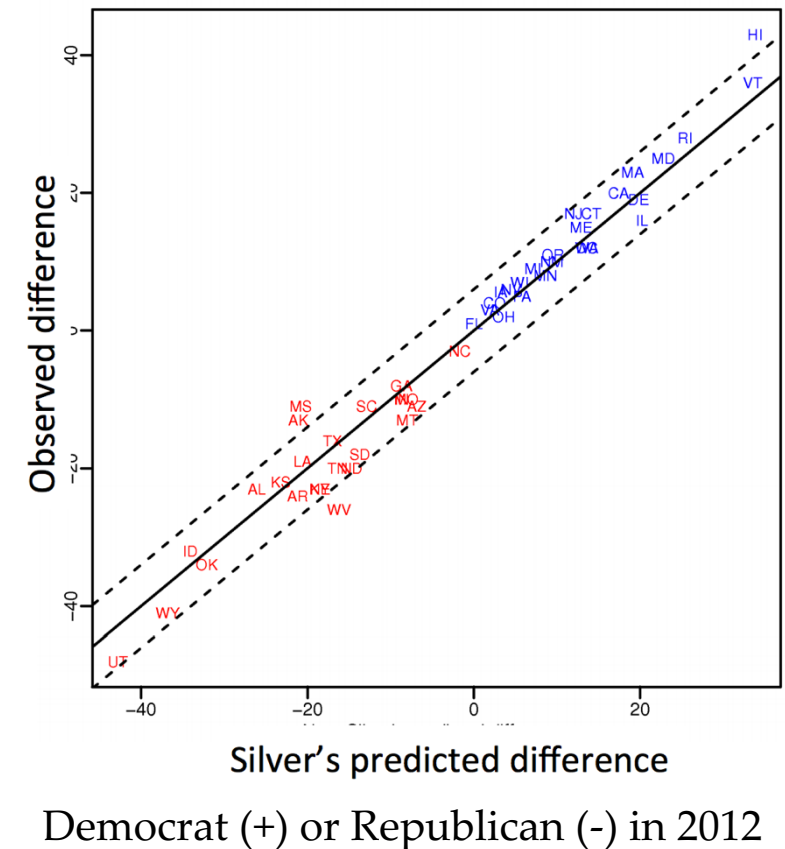
Polling: 2016

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

ELECTION DAY

Share of weight given to **fundamentals**

Share of weight given to **polls**

Days until election

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
- Waaaayyyy back in 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

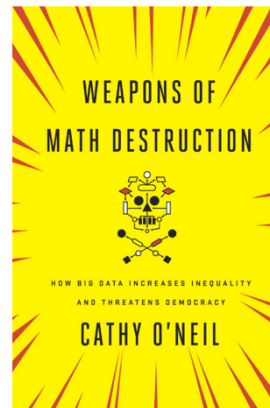http://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2

Automated Decisions of Consequence…

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

textio   DOXA

zest finance

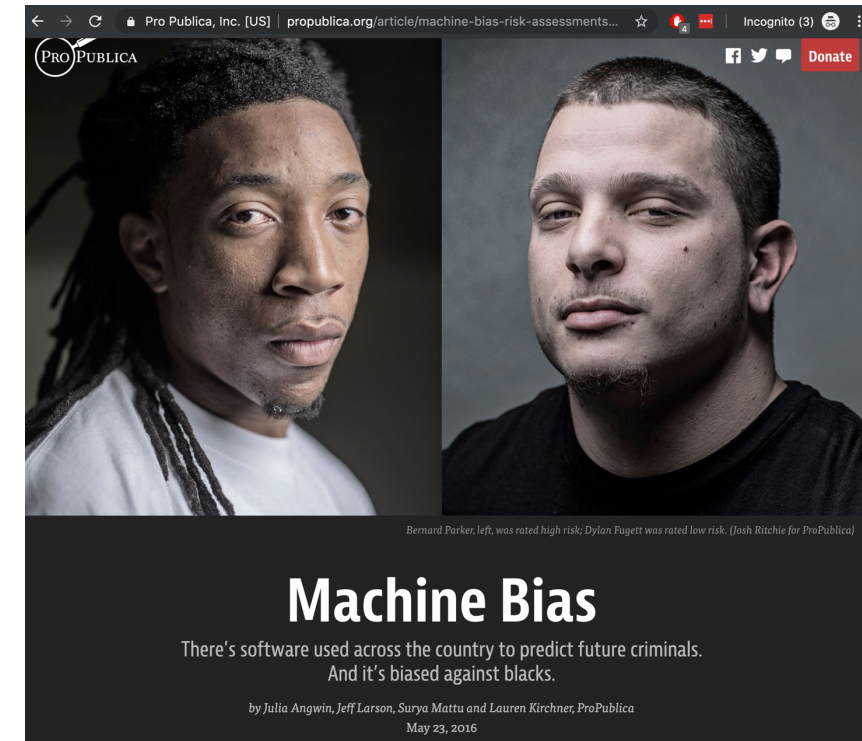LendingClub

PREDPOL

gild   GJ GapJumpers

Kabbage

NORTHPOINTE
Advancing Justice. Embracing Community.

Hiring

Lending

Policing/ sentencing

Search for minority names →
    ads for DUI/arrest records

WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

Female names →
    less freq. shown professional job opening ads

Pro Publica, Inc. [US] | propublica.org/article/machine-bias-risk-assessments... | Incognito (3)

PRO PUBLICA   Donate

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals.
And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Mechanism Design 4 Social Good

- Run by Rediet Abebe and Kira Goldner, www.md4sg.com
  - Developing Nations; Bias, Discrimination, and Fairness; Healthcare; Housing; Poverty; Labor Markets. Workshops at EC.
- **Goal:** Apply principles from mechanism design for improving access to opportunity through markets.
  - How do we setup markets to allocate resources fairly?

- Interesting talks including including Peter Romer-Friedman.
  - ACLU suit: civil rights act of 1965 says you cannot discriminate jobs, credit, or hosing based on race.
  - Facebook argued provider so not responsible.
  - https://en.wikipedia.org/wiki/Section_230_of_the_Communications_Decency_Act
  - https://www.vox.com/2019/3/21/18275746/facebook-settles-ad-discrimination-lawsuits
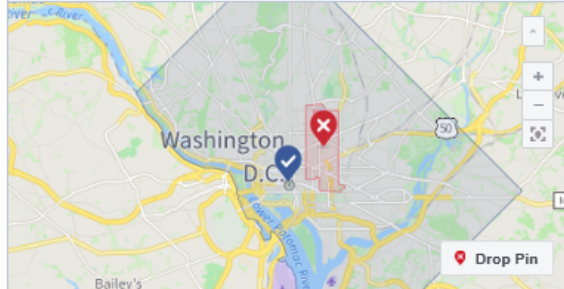
References
*Mechanism Design for Social Good.* Rediet Abebe and Kira Goldner. ACM SIGAI Matters, 2018.
Excellent work in this space by John P. Dickerson, UMD: http://jpdickerson.com/
Fairness in Deceased Organ Matching. Nicholas Mattei, Abdallah Saffidine, Toby Walsh. AIES 2018.
Mechanisms for Online Organ Matching. Nicholas Mattei, Abdallah Saffidine, Toby Walsh. IJCAI 2018.

"… 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
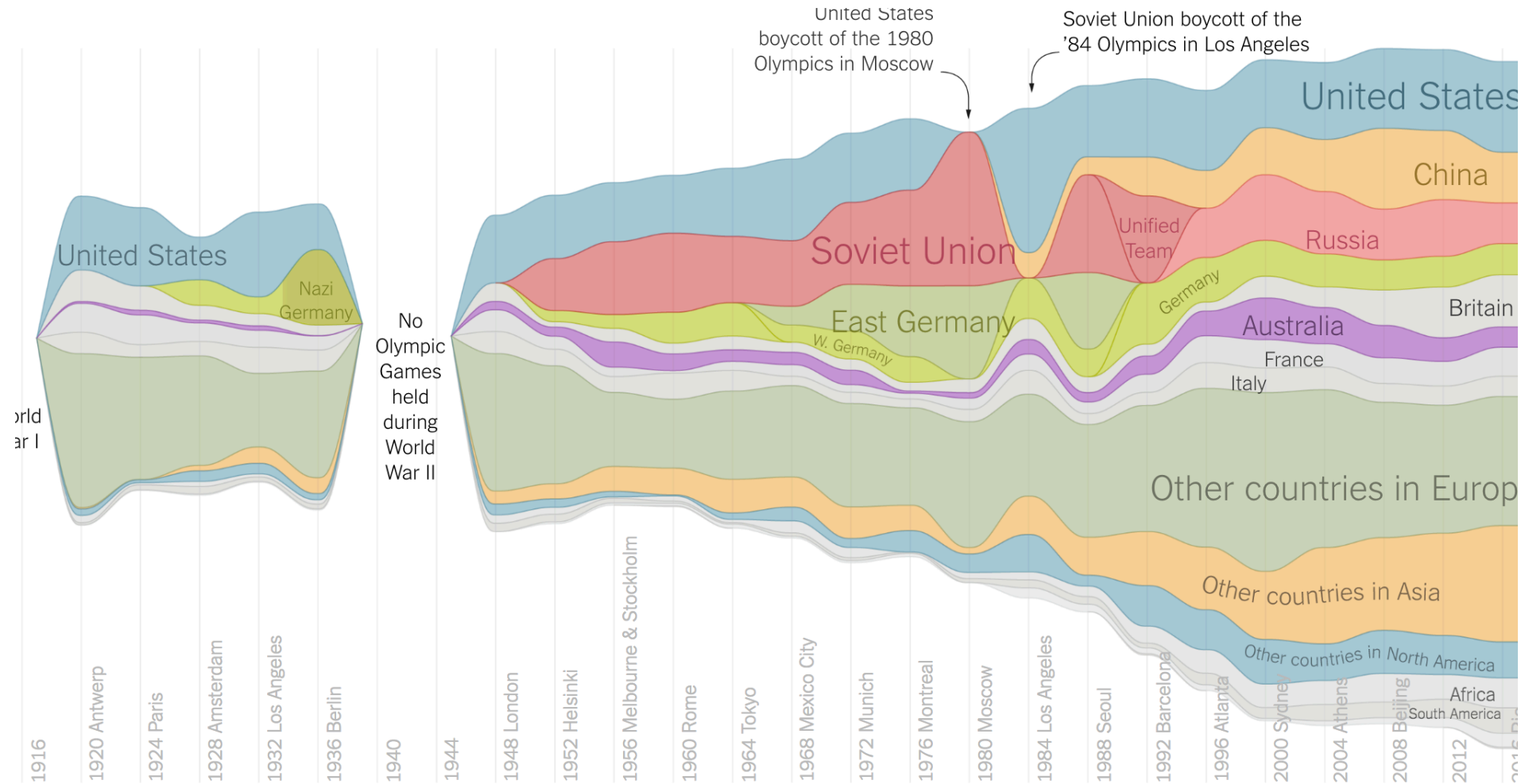
Olympic Medals Per Capita

- http://www.medalspercapita.com/#medals-per-capita:summer

MEDALS PER CAPITA
Olympic Glory in Proportion

★ ★ ★ ★   ★ ★ ★ ★

All Summer Games ▲▼

Total Medals per Capita

Gold Medals per Capita

Weighted Medals per Capita

Weighted Medals by GDP

Weighted Medals by Area

Weighted Medals by Latitude

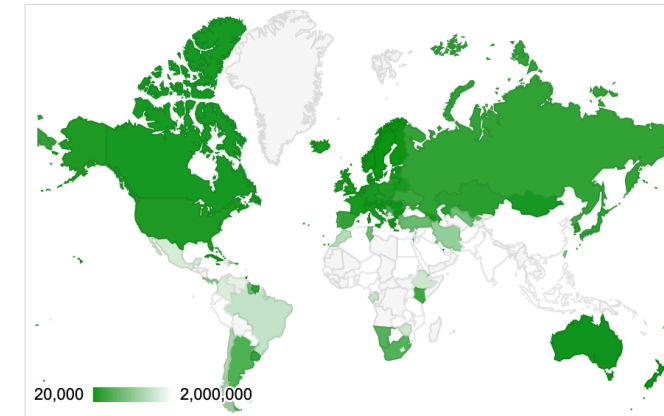Weighted Medals by Team Size

Total Medals
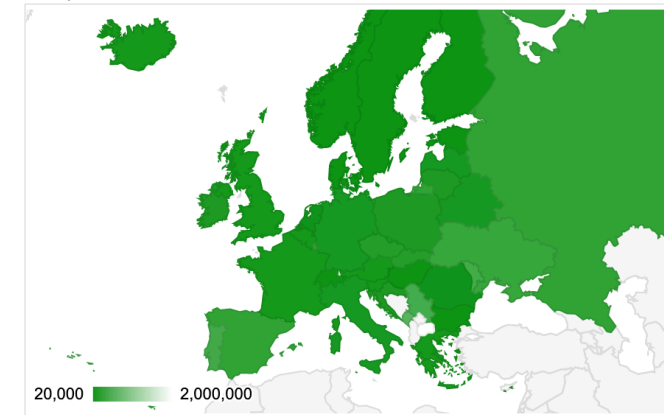
Gold Medals

Population

About

Updated February 23, 2018 at 9:51am EST

👍 Like | Share | 38K people like this. Sign Up to see what you

🐦 Tweet

🐦 Follow @MedalsPerCapita | 120 followers

Rank	Country	Medals	Population	Population per Medal
1	Finland	303	5,407,040	17,845
2	Sweden	494	9,490,683	19,211
3	Hungary	491	9,962,000	20,289
4	Bahamas	13	353,658	27,204
5	Denmark	194	5,580,516	28,765
6	Norway	153	5,005,700	32,716
7	Bulgaria	217	7,364,570	33,938
8	Jamaica	78	2,705,827	34,690
9	New Zealand	117	4,432,620	37,885
10	Estonia	34	1,318,005	38,764
11	East Germany	409	16,111,000	39,391
12	Switzerland	191	7,870,100	41,204
13	Australia	496	22,880,619	46,130
14	Cuba	219	11,241,161	51,329
15	Grenada	2	110,821	55,410
16	Netherlands	285	16,731,770	58,707
17	Romania	306	19,042,936	62,231
18	Bermuda	1	64,237	64,237
19	Trinidad and Tobago	19	1,317,714	69,353
20	Great Britain	848	62,262,000	73,422
21	Belgium	148	10,951,266	73,995
22	Iceland	4	319,575	79,893
23	Slovenia	23	2,057,540	89,458
24	France	712	65,350,000	91,783

World

20,000 — 2,000,000

Europe

20,000 — 2,000,000

Caribbean

45

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
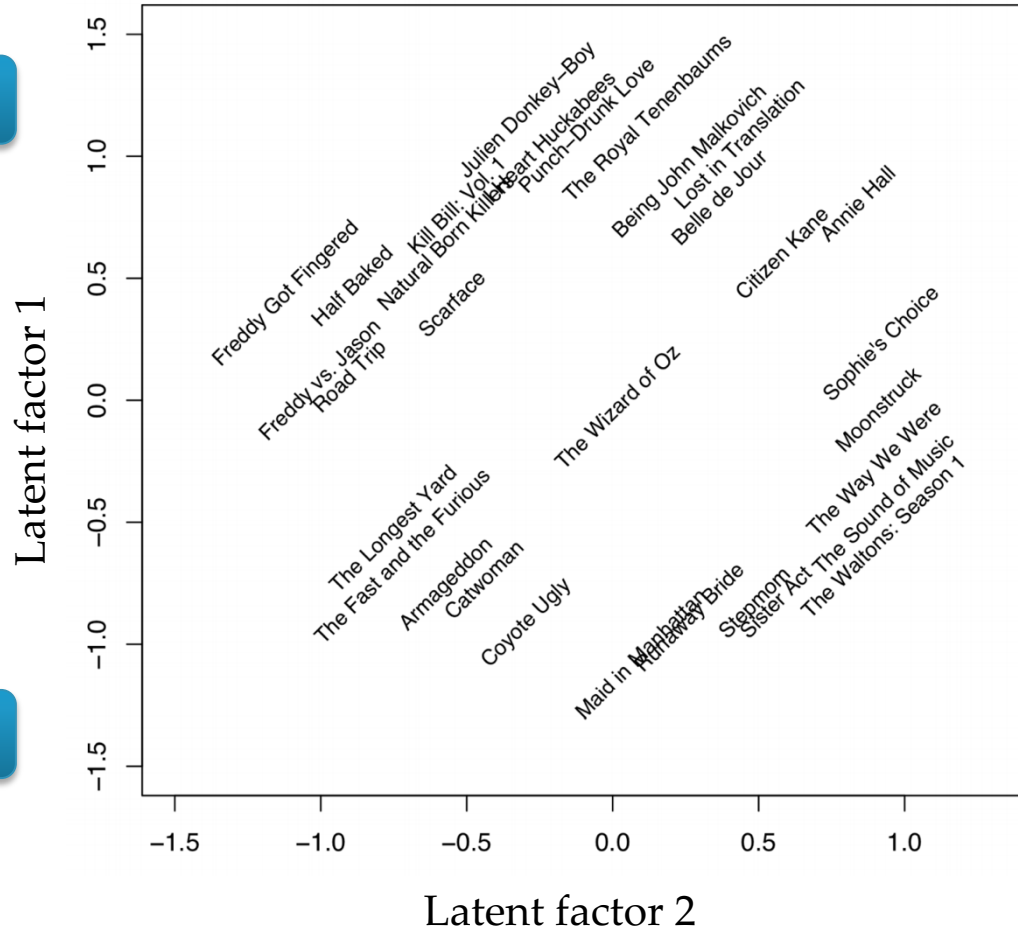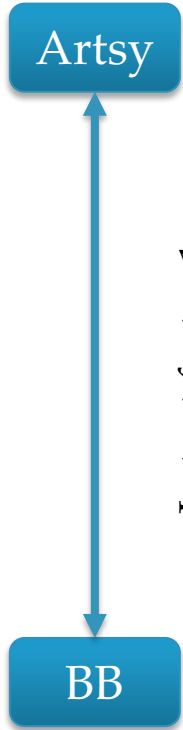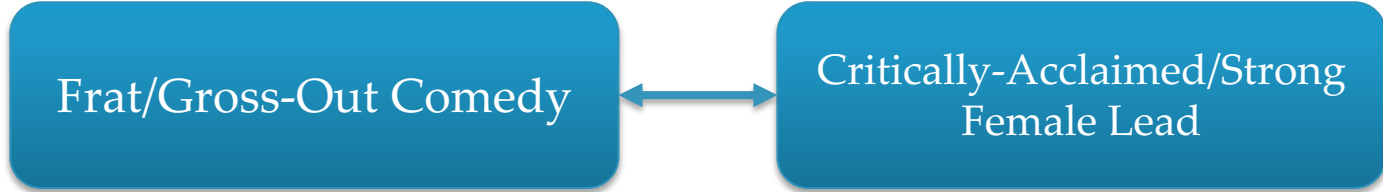
Image courtesy of Christopher Volinsky

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 (Peter Brand in the movie), data scientist *du jour*, convinces the {bad, poor} Oakland Athletics to use a quantitative aka sabermetric approach
    - (Spoiler: Red Sox offer Brand a job, he says no, they take a sabermetric approach and win the World Series.)

- DePodesta is now the head strategist for the Browns...

THE RINGER

2019 NFL PREVIEW

How the Browns and Paul DePodesta Brought Moneyball to the NFL

Two decades ago, the strategist was an oddity with the Oakland A's. Now, he's an archetype, applying the lessons of baseball's analytics revolution to football one Cleveland draft pick at a time.

By Michael Baumann | Aug 23, 2019, 6:30am EDT
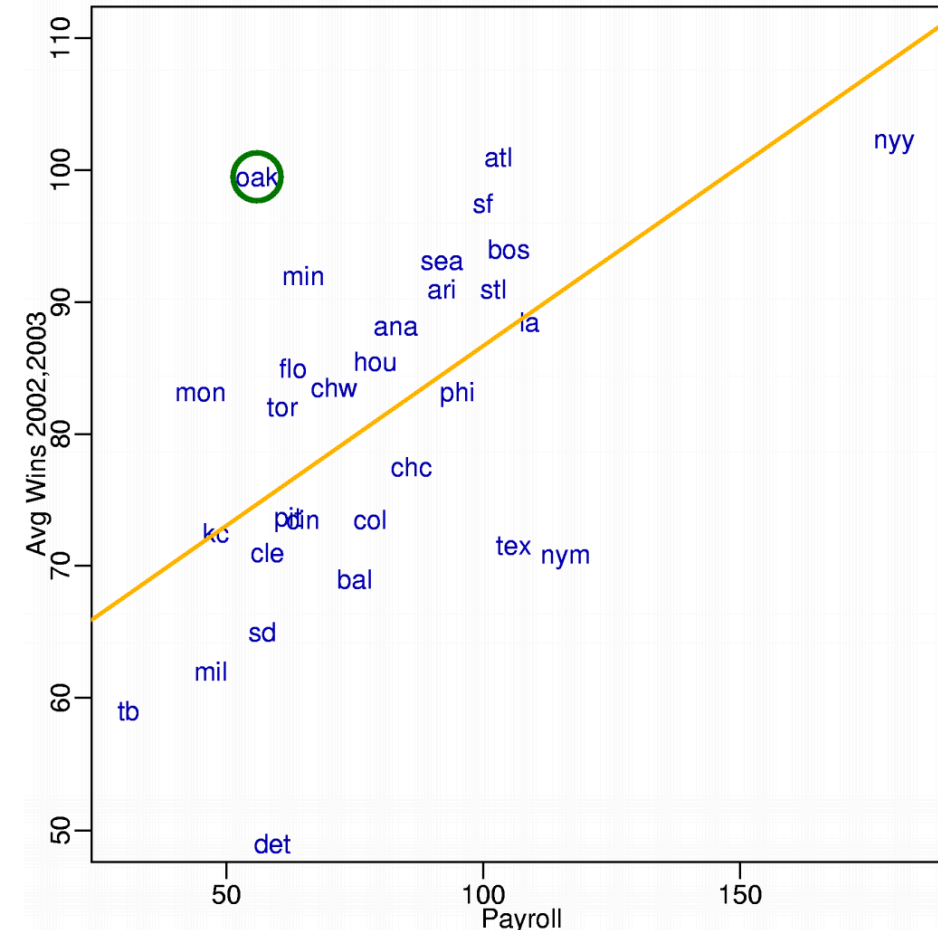
1. Data scientist

Shutterstock

Overall job score (out of 5.0): 4.8

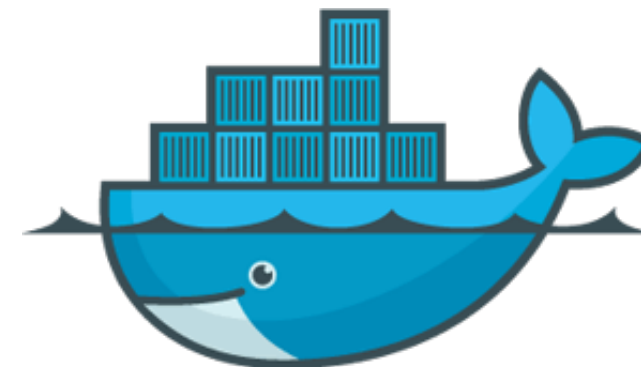Job satisfaction rating (out of 5.0): 4.4

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

- 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/