INTRODUCTION TO DATA SCIENCE

JOHN P DICKERSON

Lecture #1 – 01/26/2017

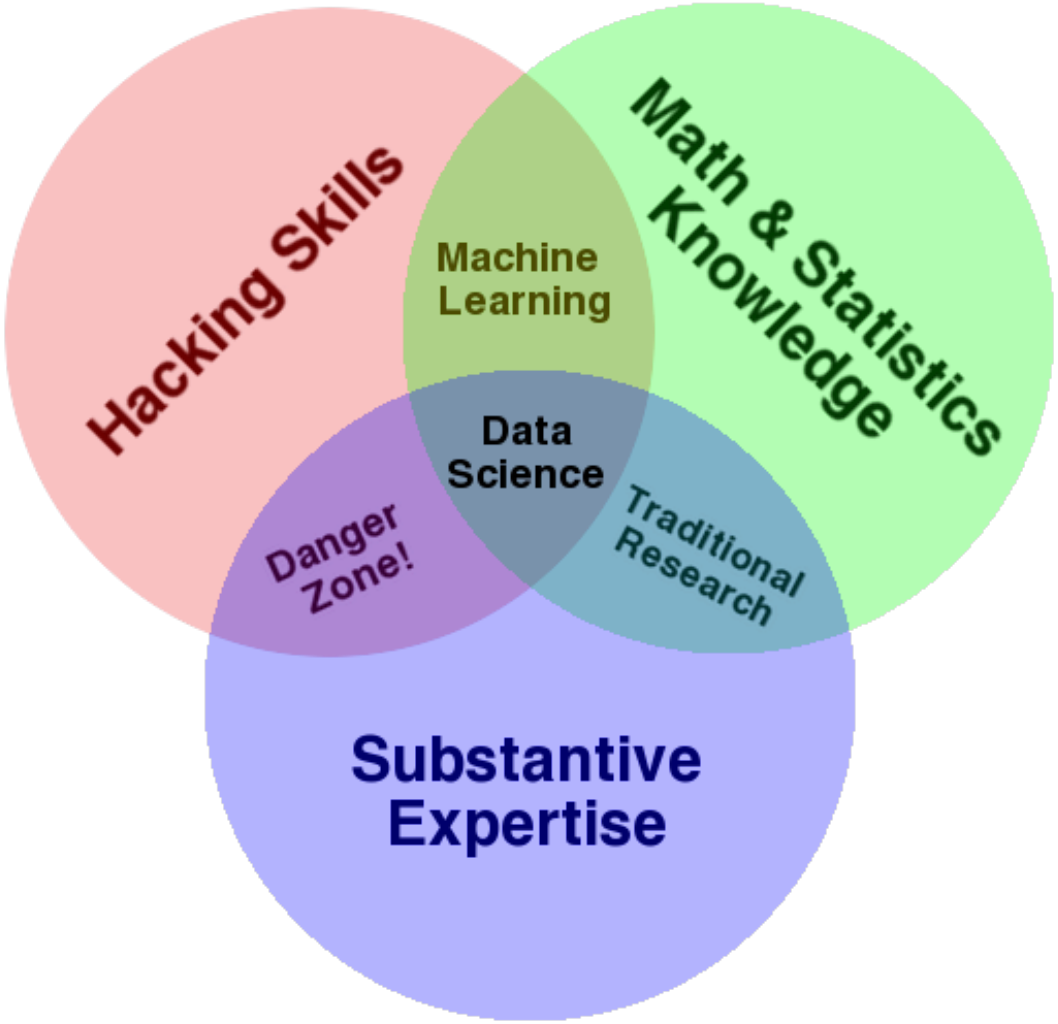
CMSC320
Tuesdays & Thursdays
3:30pm – 4:45pm

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



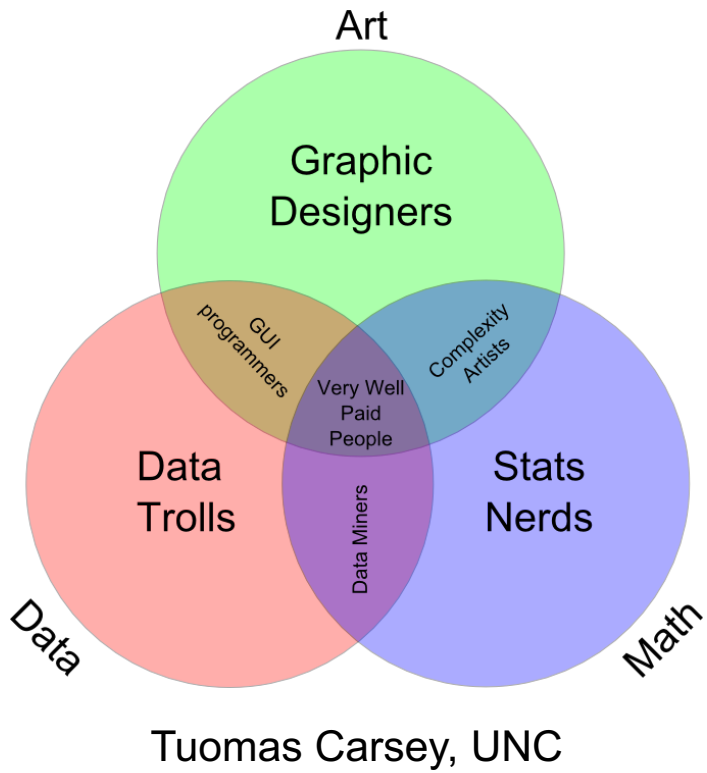
Drew Conway
CEO, Alluvium (analytics company)

MANY DEFINITIONS

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

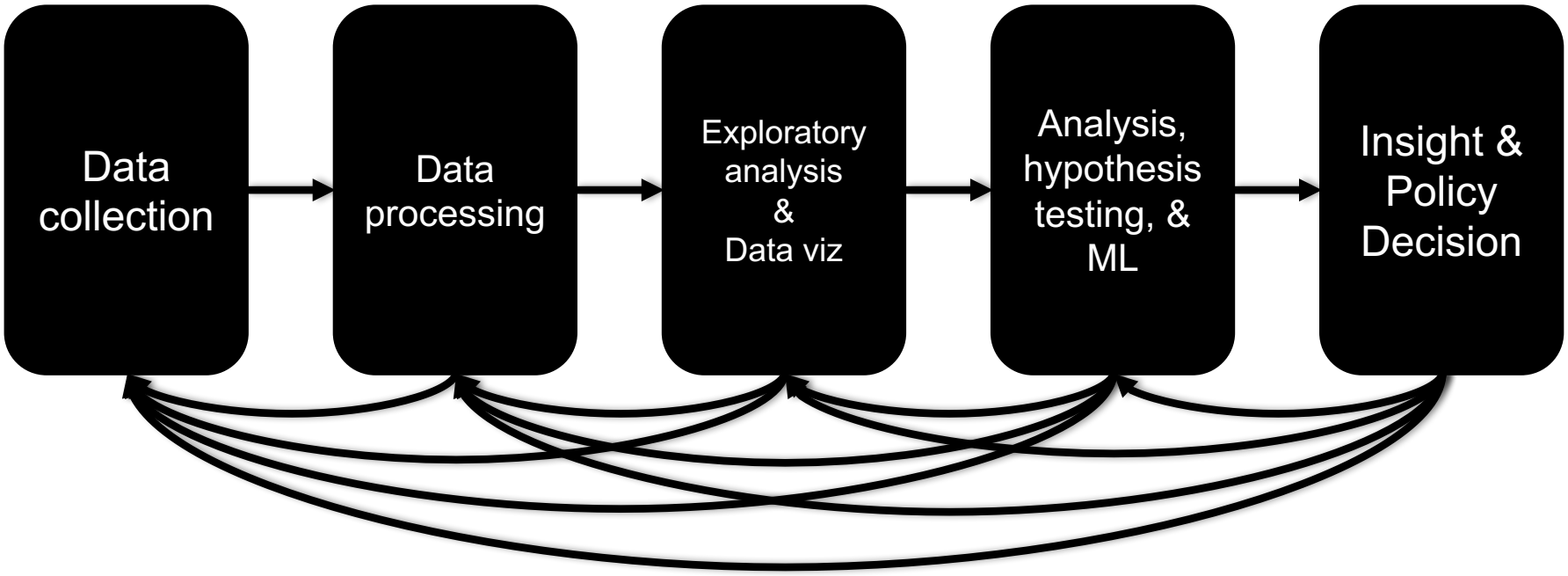
Interdisciplinary: statistics, computer science, operations research, statistical and machine learning, data warehousing, visualization, mathematics, information science, …

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, 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

THIS COURSE

You'll learn to take data:

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

Hal Varian

Info: https://www.cs.umd.edu/class/spring2017/cmsc320/

Piazza: piazza.com/umd/spring2017/cmsc320

PREREQUISITE KNOWLEDGE

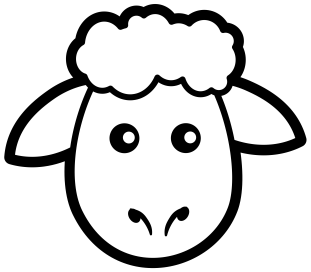
Aimed at CMSC undergrads – but likely accessible to others with programming experience and mathematical maturity.

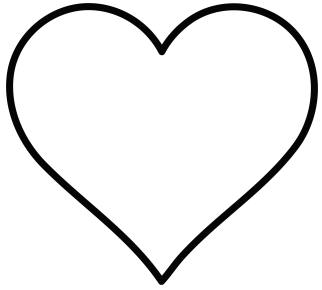
We do not assume:

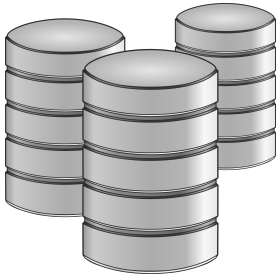
- Experience with Python, pandas, scikit-learn, matplotlib, etc …
- Deep statistics or any ML knowledge
- Database or distributed systems knowledge

We do assume:

- You want to be here!







WHO AM I?

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Carnegie Mellon University
School of Computer Science

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

http://jpdickerson.com

UNOS DONATE LIFE
UNITED NETWORK FOR ORGAN SHARING

OPTIMIZED MARKETS

IBM

THE DIAN FOSSEY GORILLA FUND INTERNATIONAL

WHO ARE YOU?

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

STAT400?
CMSC422?
CMSC424?

Register on Piazza: piazza.com/umd/spring2017/cmsc320

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

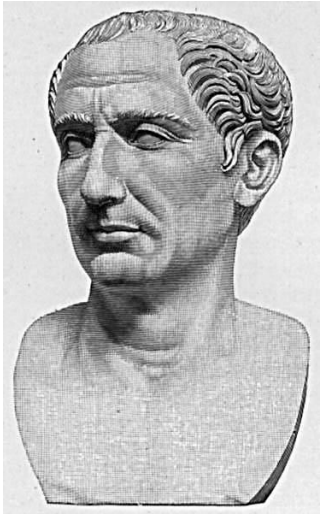
Next 9 lectures: statistical modeling & ML

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

Midterm

Final 8 lectures: advanced topics

- Dimensionality reduction, distributed learning, big data, distributed computation
- *Either* group presentations or more 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 a PDF of a Jupyter notebook on ELMS

GRADE #2: CLASS PARTICIPATION

Please please please please please do the required reading, if available, before coming to class!

Earn full credit via:

- Lecture participation
- Piazza participation
- Regular attendance at office hours (can be just to chat!)

Please I'm so lonely …

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

GRADE #3: MIDTERM

You know what this is.

Will cover roughly the first 2/3 of class:

- Qualitative (more)
- Quantitative (less)

Currently scheduled for April 13th (might change)

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.io or on UMD's servers) and will strengthen your portfolio.

GRADE BREAKDOWN

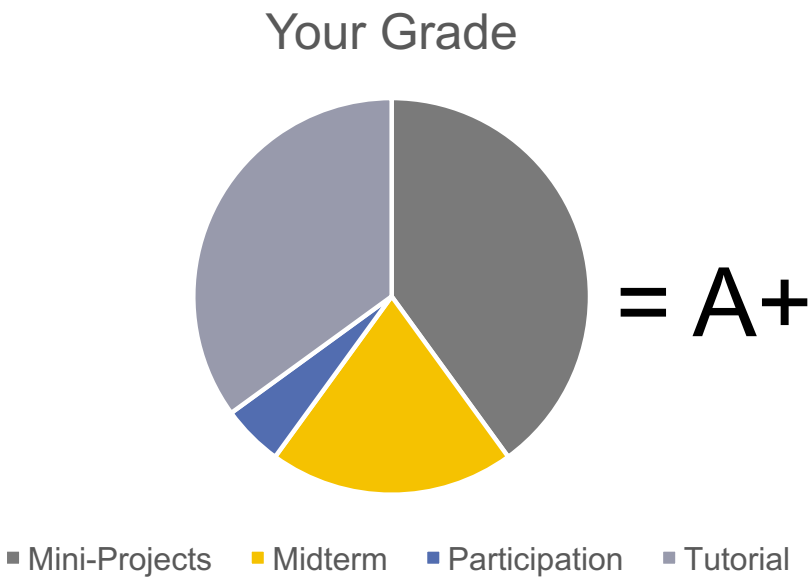
40% mini-projects:

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

20% midterm

5% course participation

35% final tutorial



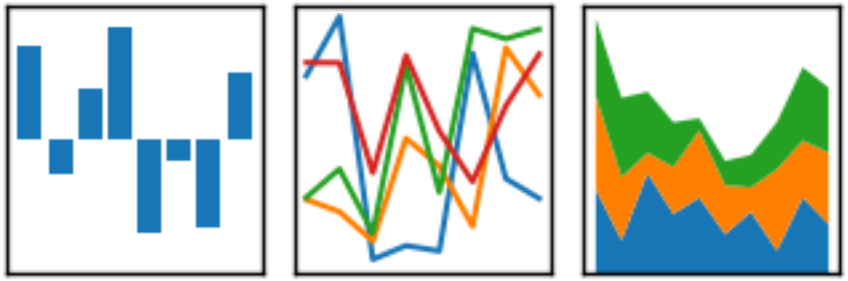
SOME TECHNOLOGIES WE WILL USE

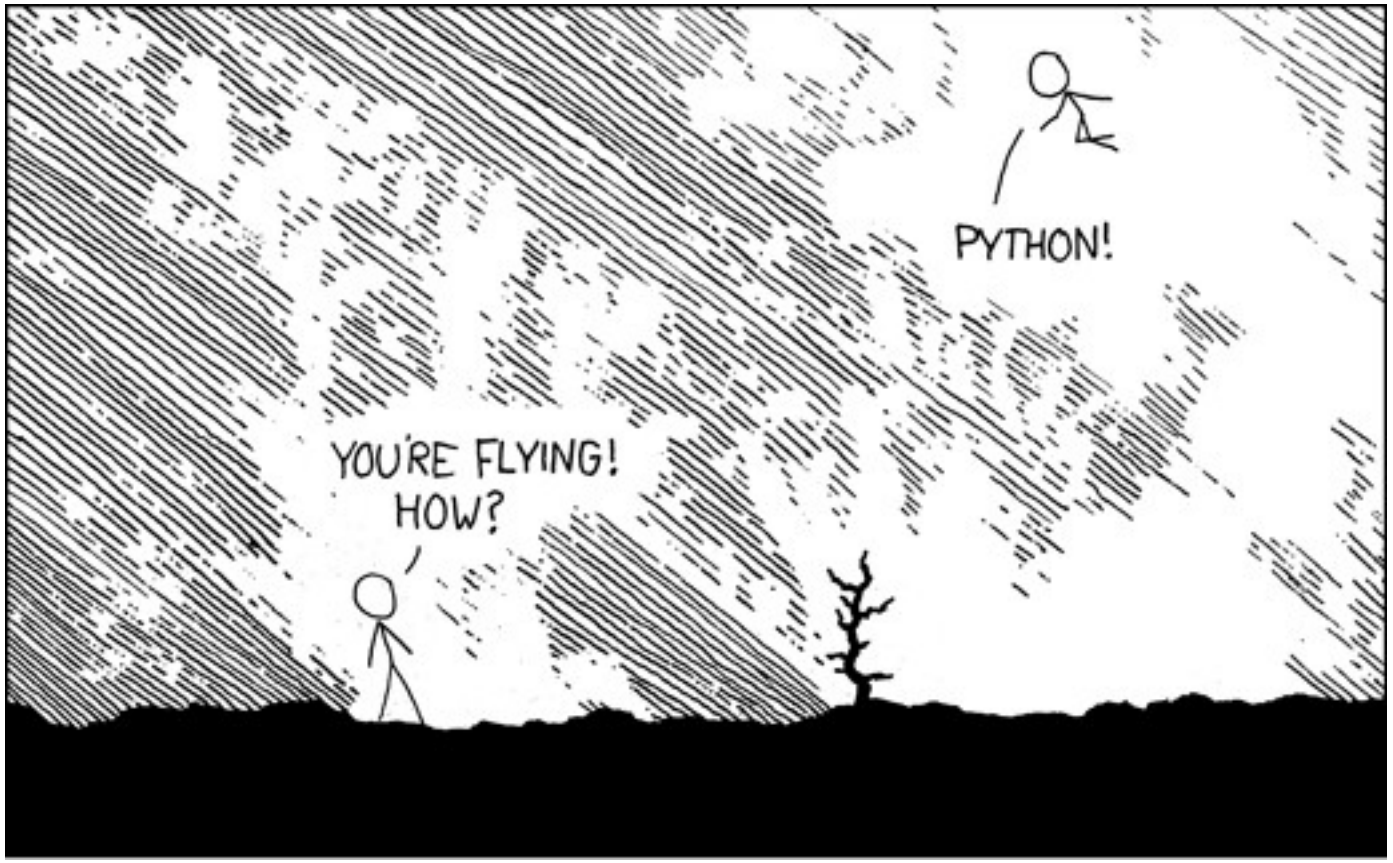
python™

ANACONDA®
Powered by Continuum Analytics

pandas

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





(Don't tell CMSC330 …)

IMPORTANT WALLS OF TEXT

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.

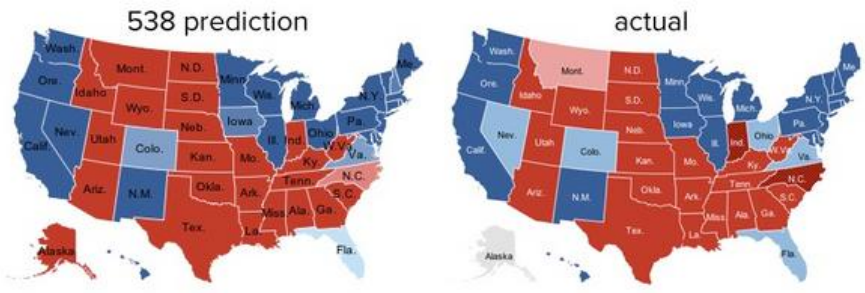
Everything you hand in must be in your own words, and based on your own understanding of the solution. If someone helps you understand the problem during a high-level discussion, you are not cheating. We strongly encourage students to help one another understand the material presented in class, in the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

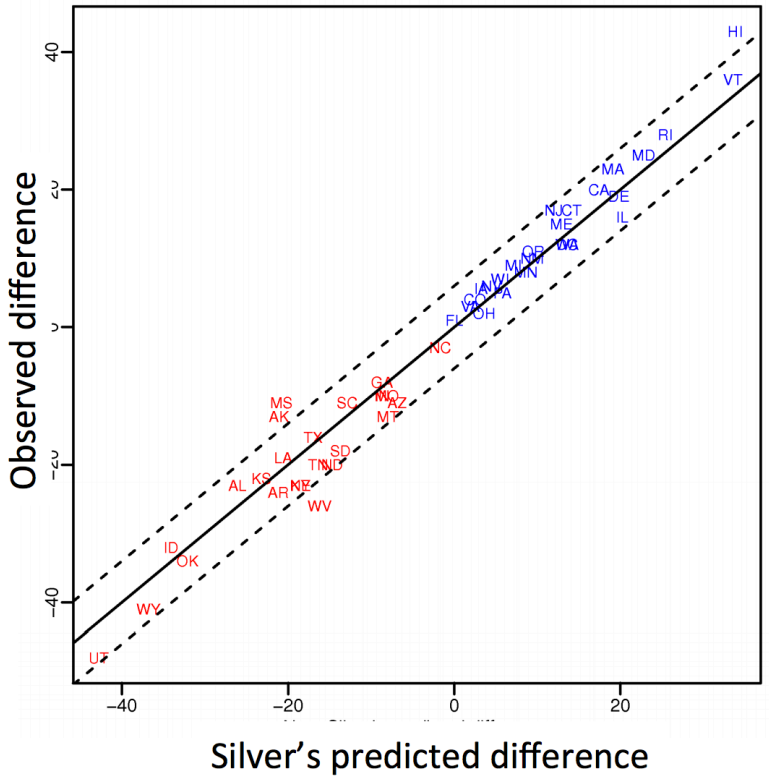
Nate Silver uses a simple idea – taking a principled approach to aggregating polling instead of relying on punditry – and:

- **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

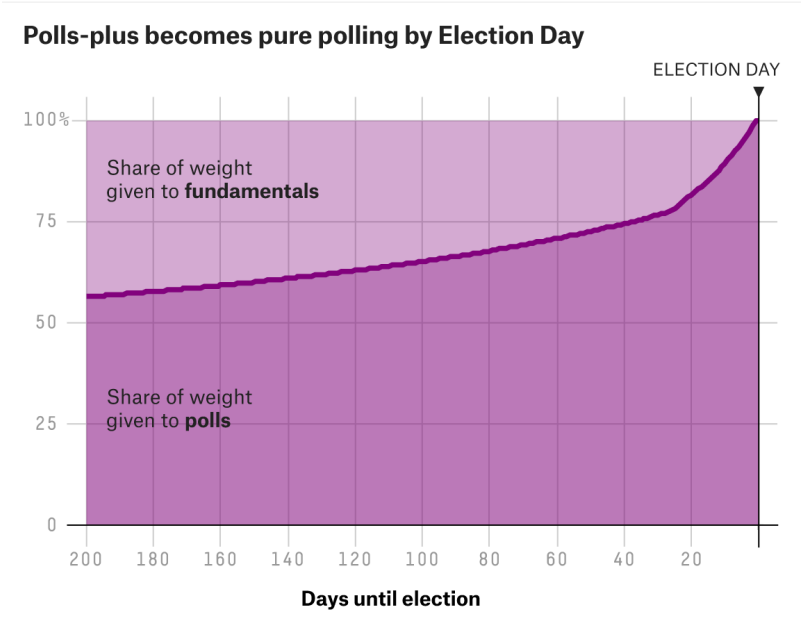
POLITICS

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/

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

gild

GJ GapJumpers

zestfinance

LendingClub

Kabbage

PREDPOL

NORTHPOINTE
Advancing Justice. Embracing Community.

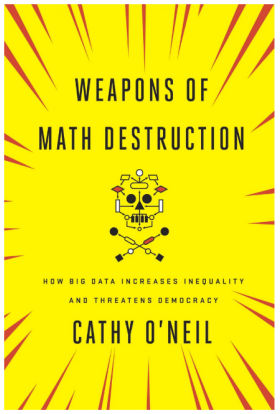
Hiring

Lending

Policing/ sentencing

Search for minority names → ads for DUI/arrest records

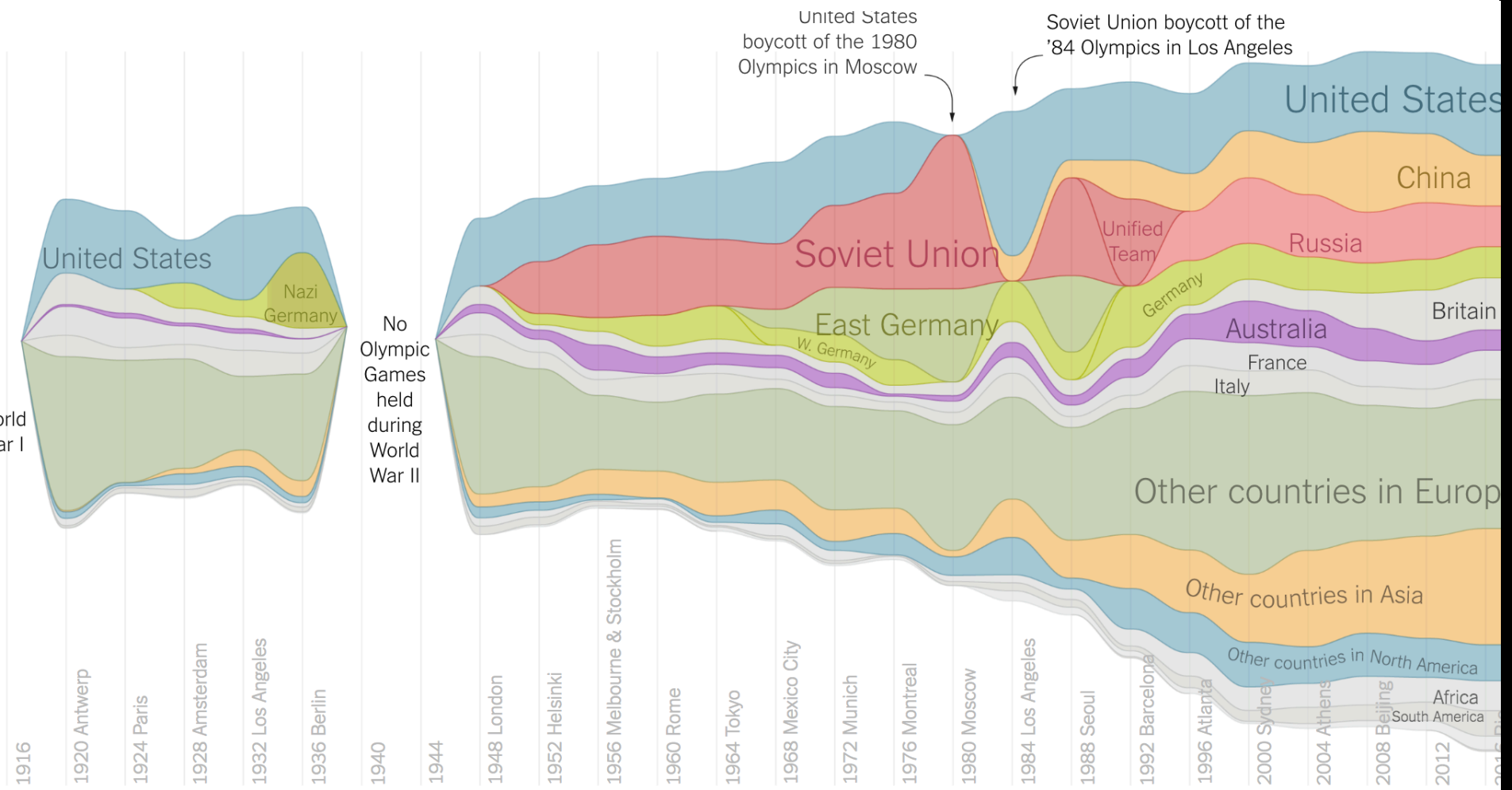
Female cookies → less freq. shown professional job opening ads



"… 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

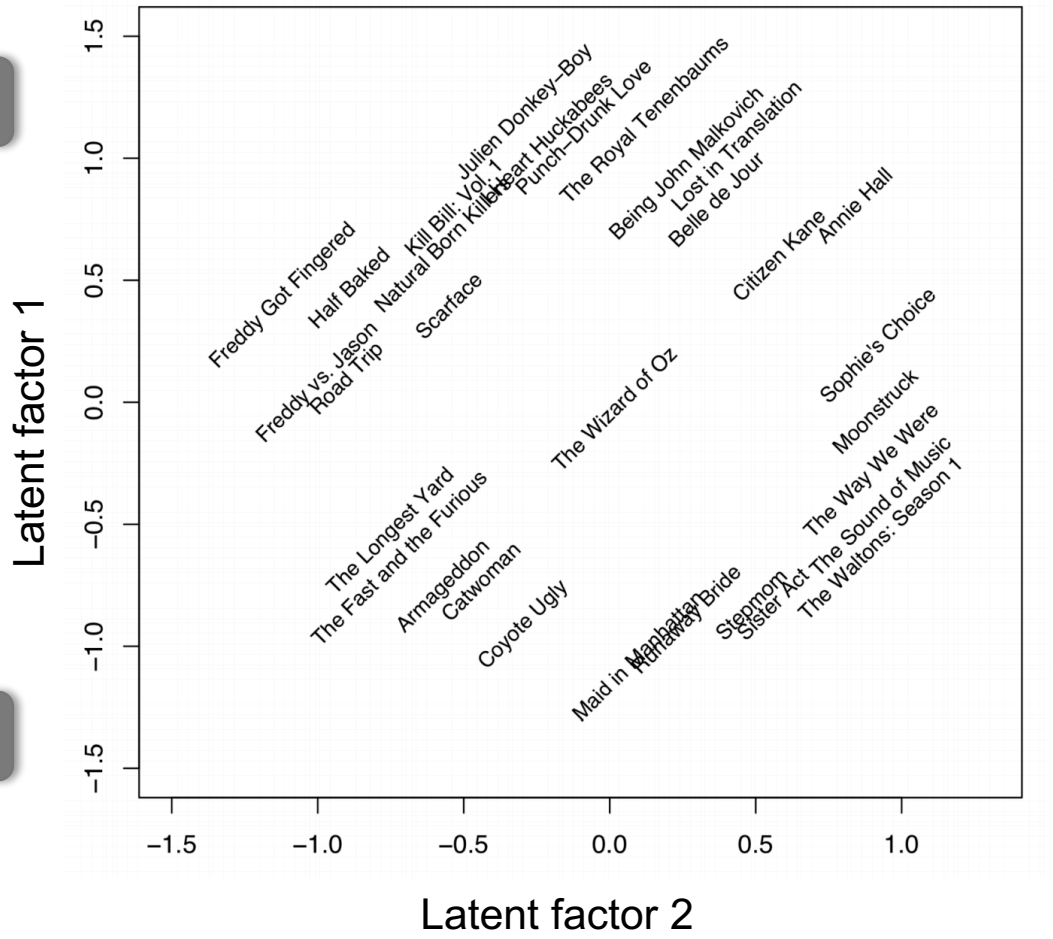
Netflix Prize: $1MM to the first team that beats our in-house engine by 10%

- **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

NETFLIX PRIZE II

Frat/Gross-Out Comedy ⟷ Critically-Acclaimed/Strong Female Lead

Artsy ↕ BB



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

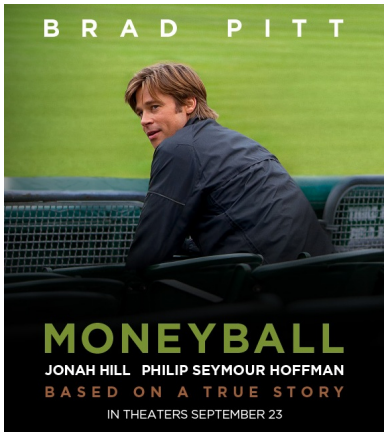
In 2007, UT Austin managed to deanonymize portions of the original released (anonymized) Netflix dataset:

- **??????????**
- **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

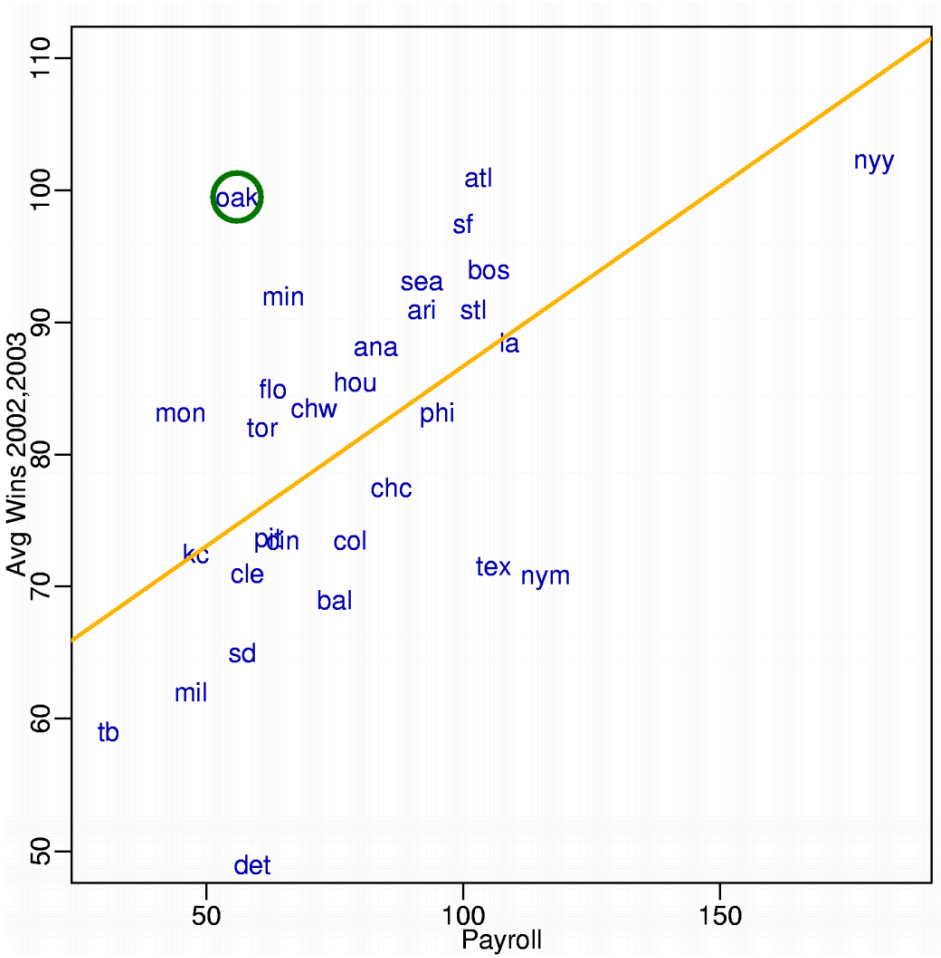
MONEYBALL



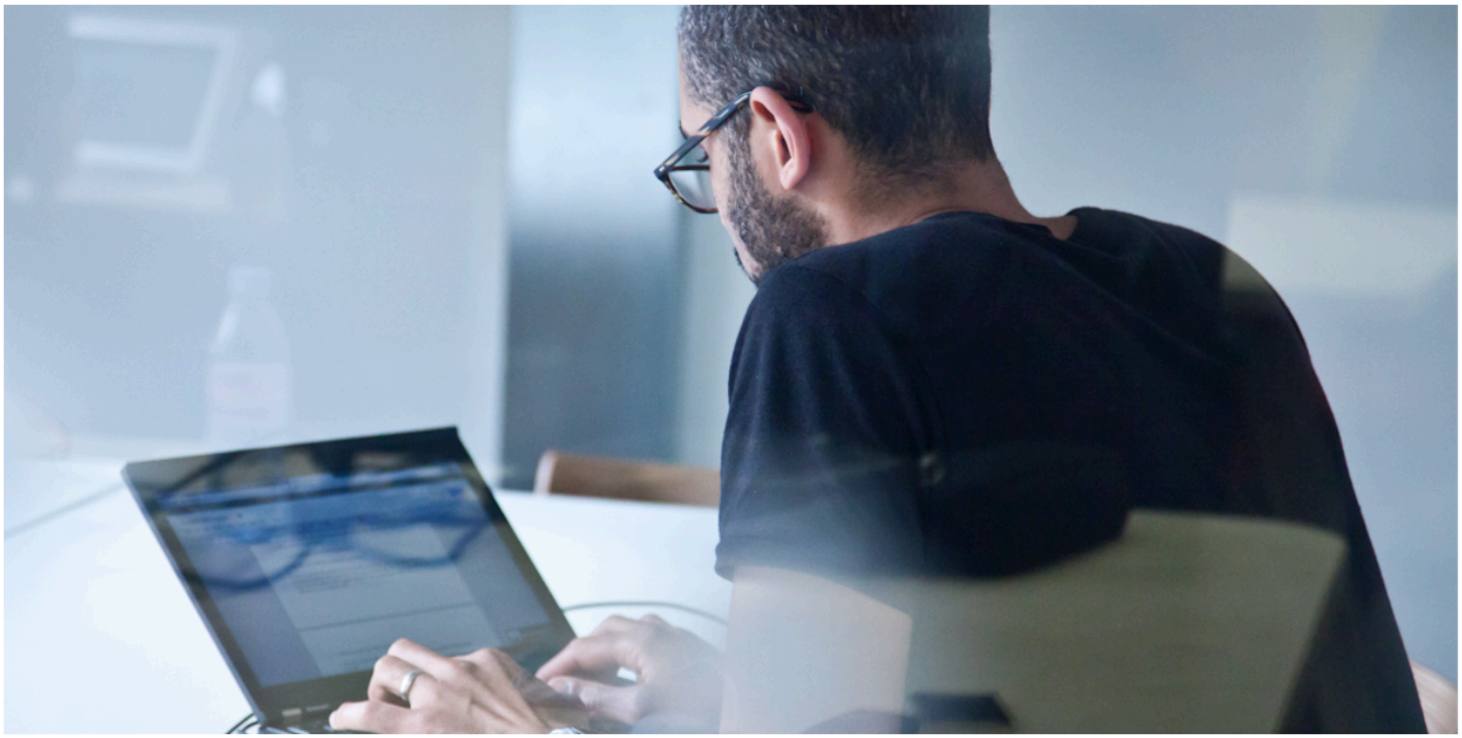
Baseball teams drafted rookie players primarily based on human scouts' opinions of their talents

Peter Brand, 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.)



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/spring2017/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

Over the weekend, install Anaconda!

- Works on *nix, OSX, Windows
- https://www.continuum.io/downloads
- Choose Python 3.5!

ANACONDA

Powered by Continuum Analytics®

NEXT CLASS:

SCRAPING DATA WITH PYTHON