Introduction

Everything Data

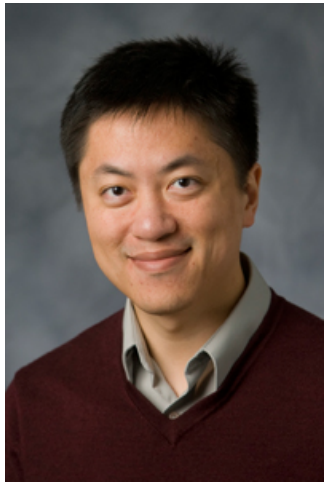
CompSci 216 Spring 2015

DUKE
COMPUTER SCIENCE

About us: instructors

Ashwin Machanavajjhala:
data privacy, massive data analytics

Jun Yang: data-intensive
systems, computational journalism

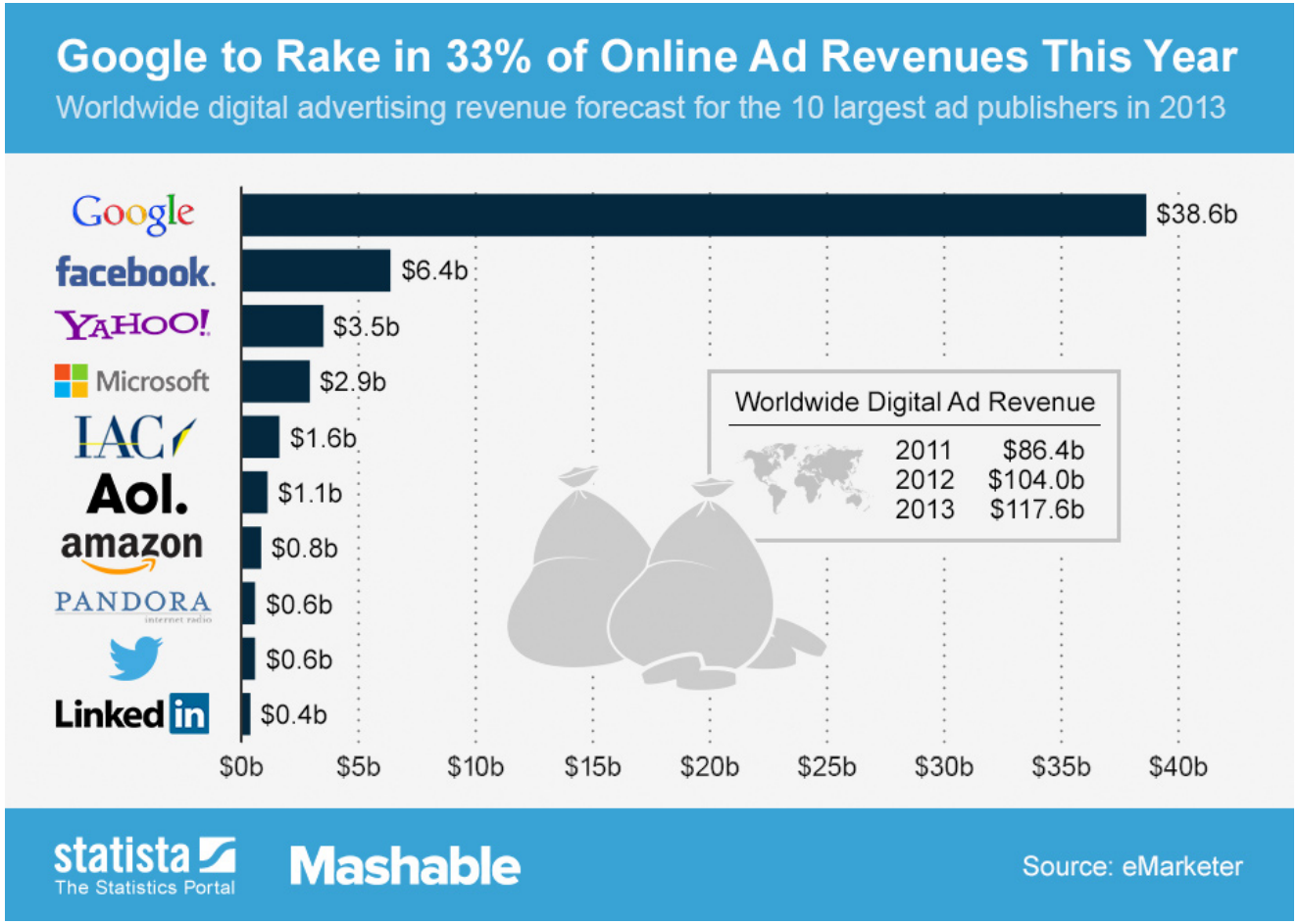


About us: TAs

Prajakta Kalmegh (Grad TA):
optimizing big-data analytical workloads

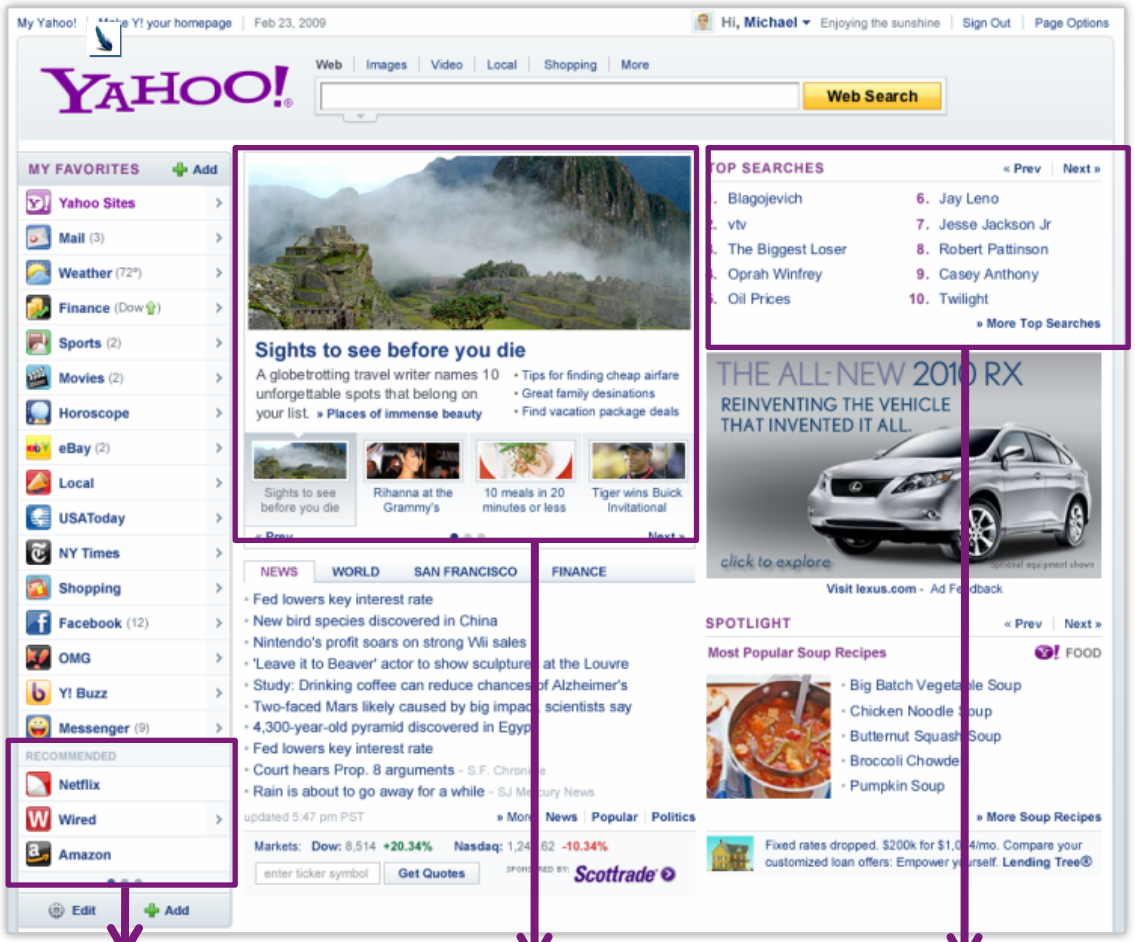
More UTAs are still being recruited…

Let's talk about $$$



In perspective: 87% of Google's revenue comes from online ads (as of 2012)

Data and business



Recommended links

+79% clicks

vs. randomly selected

Personalized News Interests

+250% clicks

vs. editorial one-size-fits-all

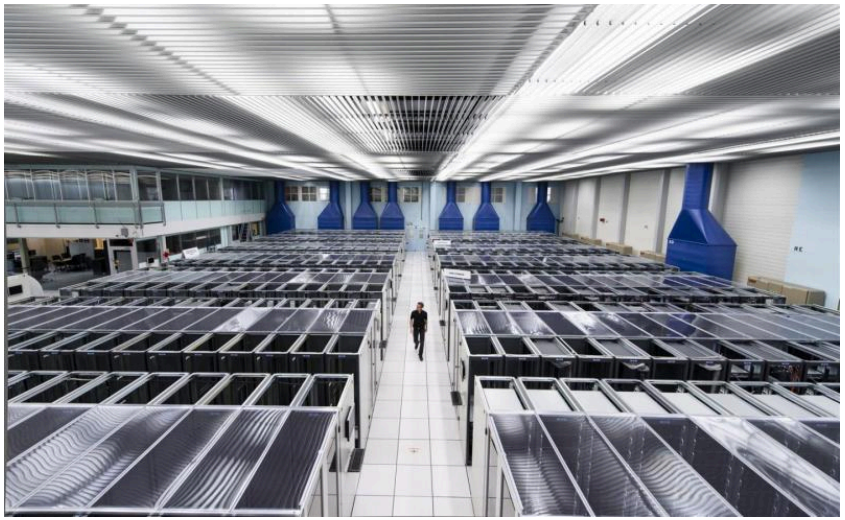
Top Searches

+43% clicks

vs. editor selected

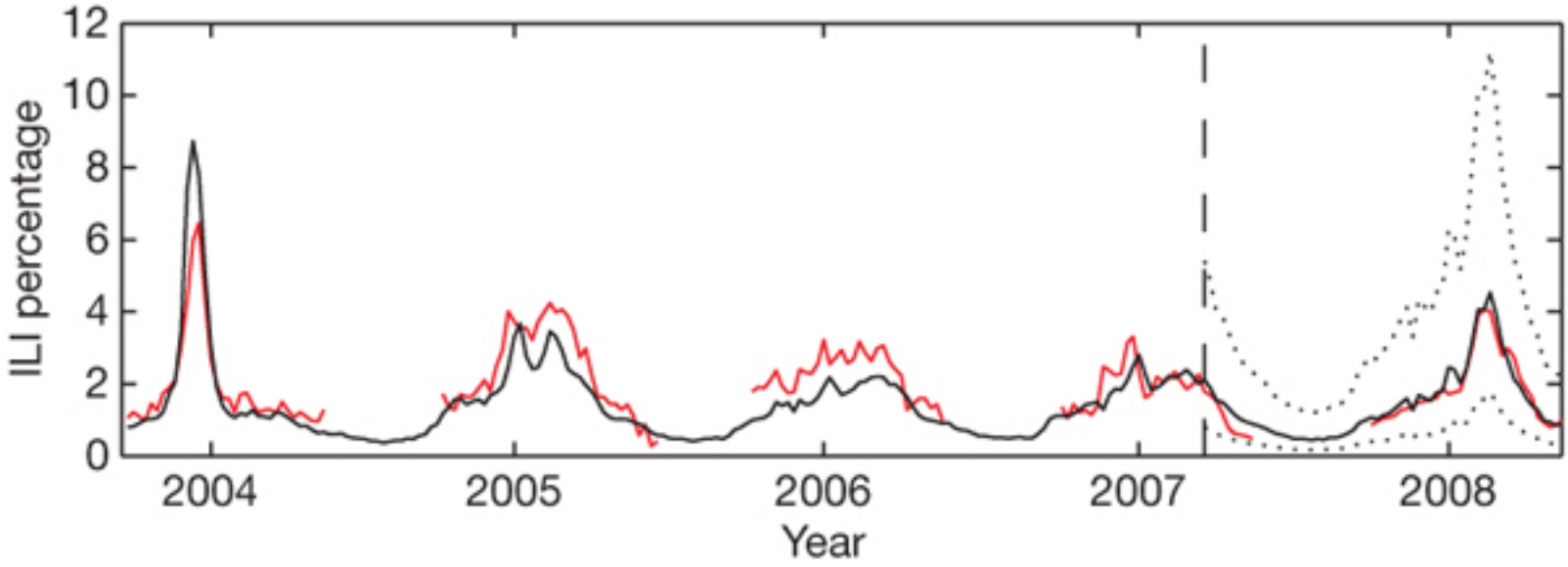
Data and science

- The world's largest particle collider at CERN—where the Higgs boson was confirmed—generates 30 petabyte of data per year
- CERN's data center has 11,000 servers with 100,000 cores... yet it still can't crunch all data!



http://home.web.cern.ch/about/computing

Data and health

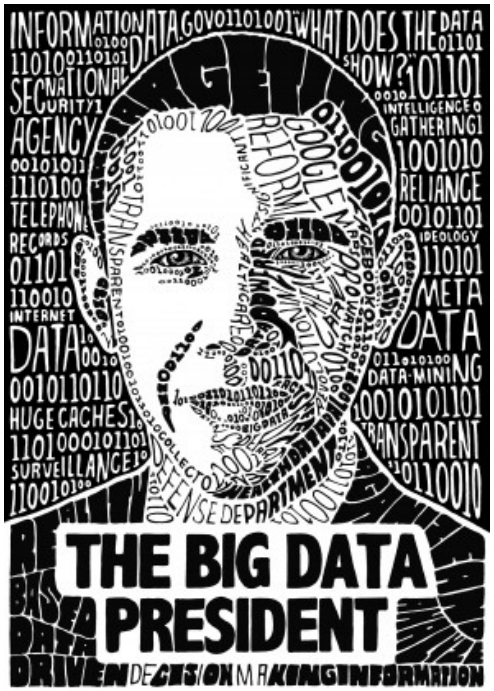


Red: official numbers from Center for Disease Control and Prevention; weekly
Black: based on Google search logs; daily (potentially instantaneously)

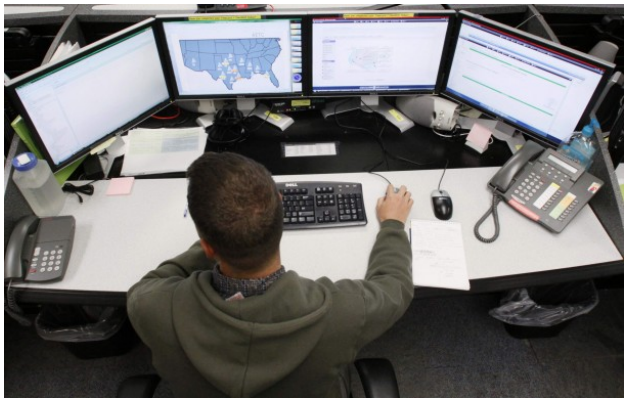
Detecting influenza epidemics using search engine query data

http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html

Data and government



http://www.washingtonpost.com/opinions/**obama-the-big-data-president**/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html



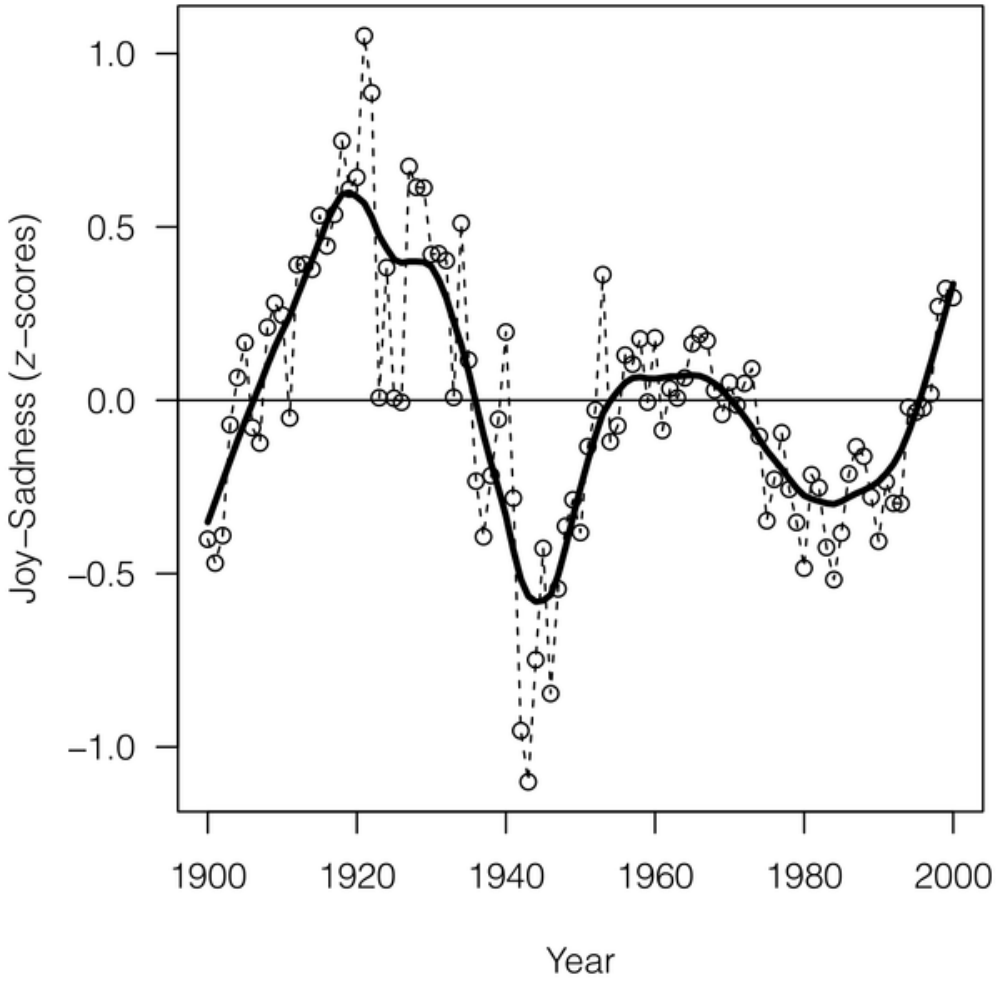
http://www.washingtonpost.com/business/economy/**democrats-push-to-redeploy-obamas-voter-database**/2012/11/20/d14793a4-2e83-11e2-89d4-040c9330702a_story.html

DATA.GOV
EMPOWERING PEOPLE

http://www.whitehouse.gov/blog/**Democratizing-Data**

http://www.theguardian.com/world/2013/jun/23/**edward-snowden-nsa-files-timeline**

Data and culture

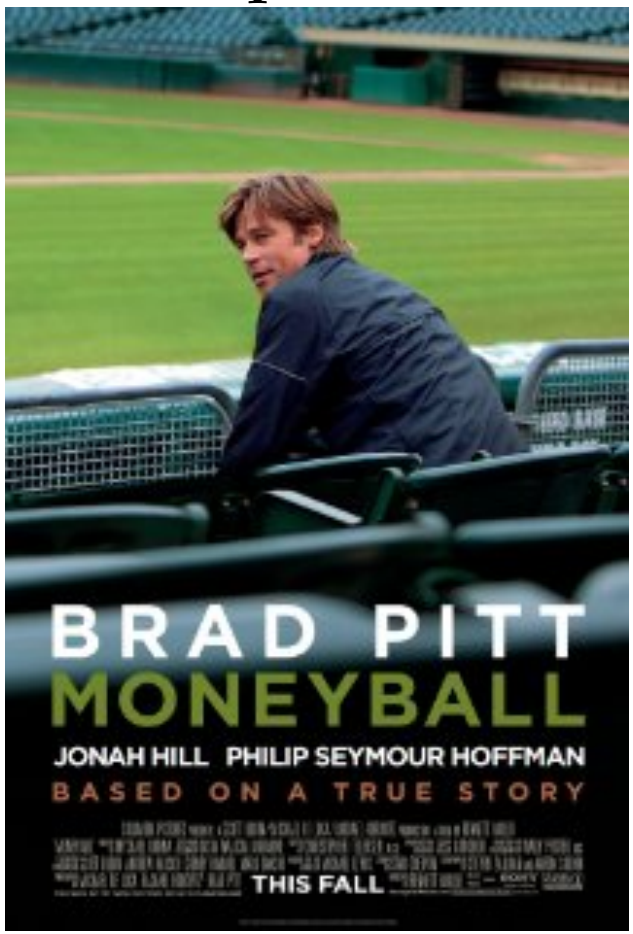


- Word frequencies in English-language books in Google's database

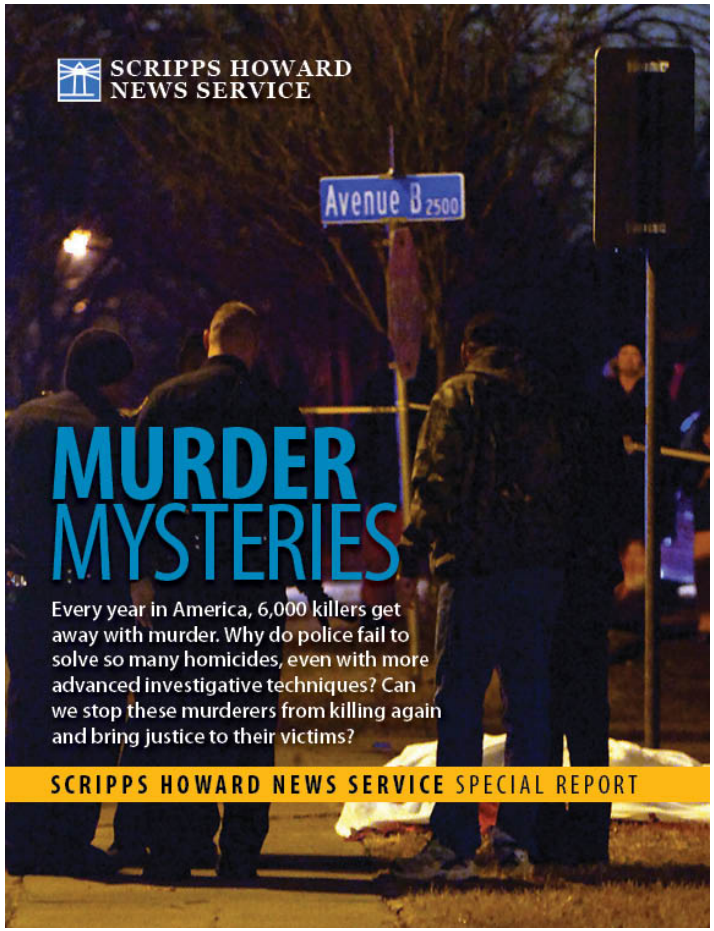
http://blogs.plos.org/everyone/2013/03/20/**what-are-you-in-the-mood-for-emotional-trends-in-20th-century-books**/

Data and ____ ☜ your favorite subject

Sports



Journalism



Hal Varian Chief Economist, Google

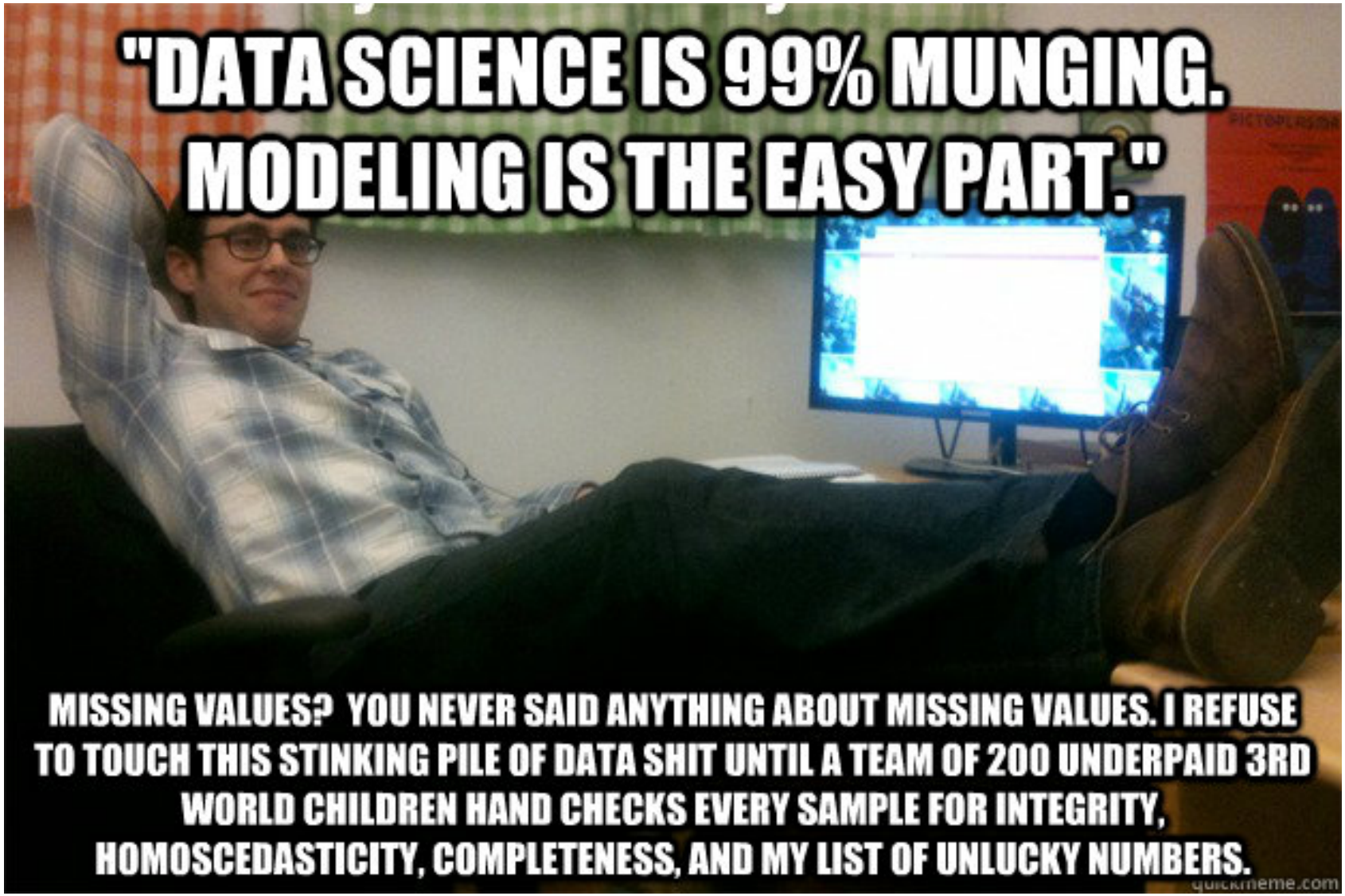
I keep saying ***the sexy job in the next ten years will be statisticians.***
People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take ***data*** *—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades…*

- Jan. 2009. http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers

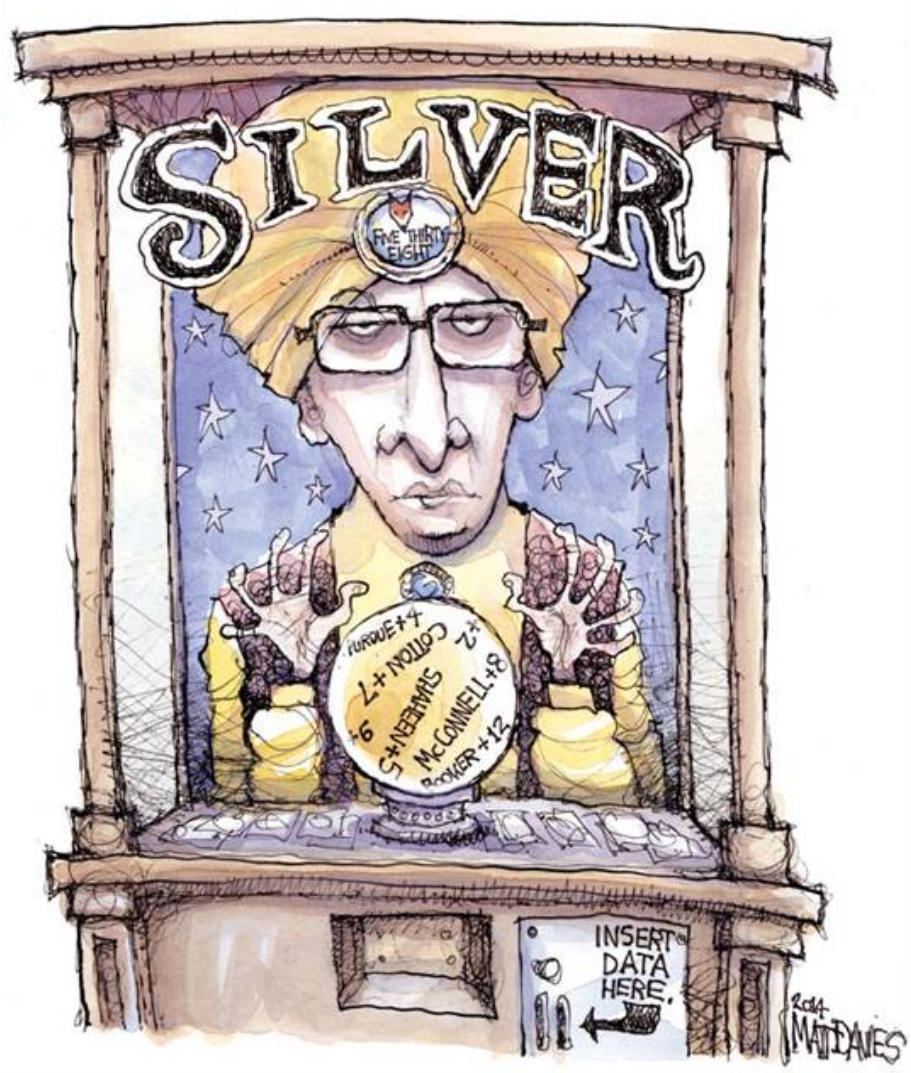
How to extract value from data

- **Wrangle** data
  - Get the data you want into the form you need for analysis
- **Analyze** data
  - Explore, query, run models, visualize…
- **Communicate** your results
  - Tell a story
  - Empower others

Data wrangling/munging



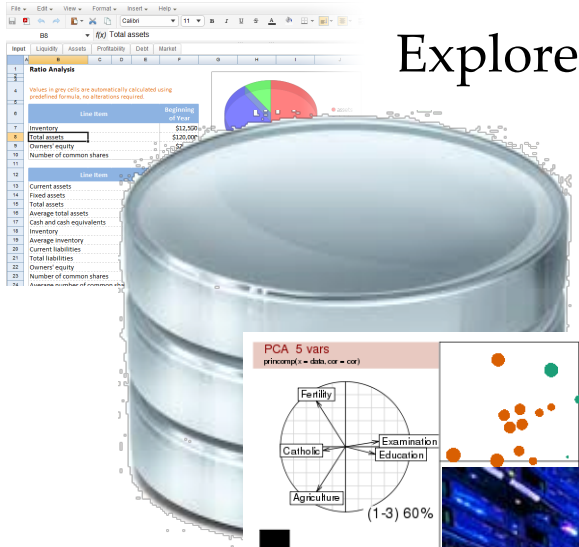
For Nate Silver…



- 70% of the time is spent on getting and cleaning data
- 15% on modeling
- 15% on programming

Personal communication, Nov. 22, 2014

Data analysis

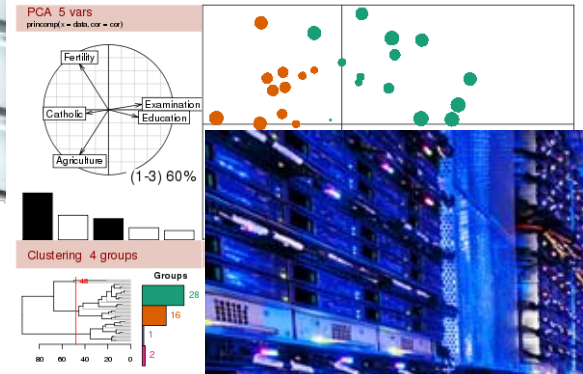


Explore and visualize, e.g., using a spreadsheet

Query, e.g., using database systems

"80% of analytics is sums and averages."

– Aaron Kimball, wibidata



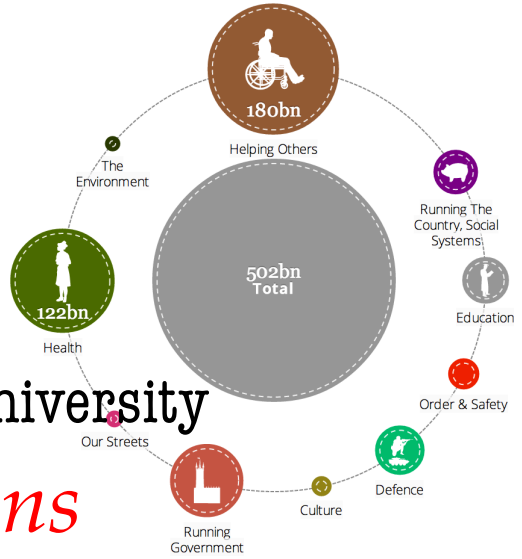
Model, detect, and predict, e.g., using R

Scale up, e.g., using MapReduce

Communicating results

"The British government spends £13 billion a year on universities."

- So?
- Try instead
  http://wheredoesmymoneygo.org/bubbletree-map.html#/~/total/education/university



"On average, 1 in every 15 Europeans is totally illiterate."

- True
- But about 1 in every 14 is under 7 years old!

http://datajournalismhandbook.org/1.0/en/understanding_data_0.html

To finish what Varian said…

I think statisticians are part of it, but it's just a part. You also want to be able to ***visualize*** *the data,* ***communicate*** *the data, and* ***utilize*** *it effectively. But I do think those skills—of being able to* ***access***, ***understand***, *and* ***communicate*** *the insights you get from data analysis—are going to be extremely important*

Pitfalls

- Hard to get right
  - Rhine Paradox of extrasensory perception
    http://lastinggems.wordpress.com/tag/rhine-paradox/
  - How accurate is Google Flu Trends?
    http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html
- Easy to abuse
  - Everyone's got a right to their own opinion
    http://www.washingtonpost.com/posteverything/wp/2014/10/13/when-it-comes-to-trickle-down-economics-everyones-got-a-right-to-their-own-opinion-but-not-their-own-facts/
  - Facebook's mood manipulation experiment
    http://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

The dark side of the force…

39% of the experts agree…

Thanks to many changes, including the building of "the Internet of Things," human and machine analysis of ***Big Data will cause more problems than it solves*** *by 2020. The existence of huge data sets for analysis will* ***engender false confidence in our predictive powers*** *and will lead many to make* ***significant and hurtful mistakes****. Moreover, analysis of Big Data will be* ***misused by powerful people and institutions with selfish agendas*** *who manipulate findings to make the case for what they want. And the advent of Big Data has a harmful impact because it* ***serves the majority (at times inaccurately) while diminishing the minority*** *and ignoring important outliers. Overall, the rise of Big Data is a big negative for society in nearly all respects.*

— 2012 Pew Research Center Report

http://pewinternet.org/Reports/2012/Future-of-Big-Data/Overview.aspx

But it's here, now!

What skills do you need?

- Domain expertise
  - Formulating problem
  - Interpreting and communicating results
- Statistics and math
  - Developing/applying quantitative models and methods to analyze data
- Computer science
  - Munging data
  - Presenting data and results
  - Developing/applying computational techniques to analyze more data faster and cheaper

Why this course?

- No single course at Duke gave you the overall picture—we want to fix that!
- With this course, we hope you will
  - Develop a holistic, interdisciplinary picture of how to deal with data
  - View data and results with a critical eye
  - Learn enough basic building blocks to go from raw data all the way to insights
  - Know what additional expertise you need for tackling bigger, harder problems

Course material

- Data wrangling
- Working with different types of data
  - Text, tabular, graph
- Working with "big" data
  - MapReduce
- Statistics
- Machine learning
  - Clustering, classification, etc.
- Visualization
- Ethics and privacy

(not necessarily in this order)

Course format

- Meetings alternate between **lectures** and hands-on, team-based **labs**
  - With **weekly homework exercises** in between
- **No exams**
- Capstone team **project**
  - Open-ended: you propose what dataset(s) you want to "take all the way"
  - Present your projects to the class at a mini-conference when semester ends

Everything Data

Team 3NN
He Jiang
Xu Han
Ying Guo

Homeownership & Housing Bubble Identification

3NN

Retweet!

TwitLab

Sign in with Twitter

Ducks

Subreddit Recommender

Nick Gordon, Howard Chung, Duke Kim

HAPPINESS

Ann Niou, Eric Wu, Kevin Wu

THE VETS: PRIMETIME

Quan Stevenson
Jordan Elkins
Chad Coviel

NFL SCOUTING COMBINE

Team Karmalytics

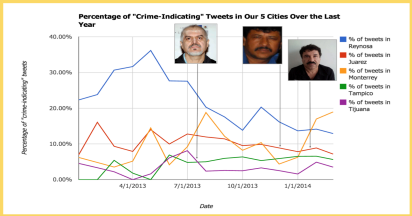
Lore's Economists

Ookillem

Predicting the popularity of a movie based on its key characteristics

Ari Adler, Zanele Munyikwa, Gary Sheng

All Cities DEA



The Decline of Drug Cartel Leaders

8/2013:
Tijuana's Eduardo Arellano-Felix is imprisoned

12/2013:
Juarez's Gonzalo Inunzo is killed

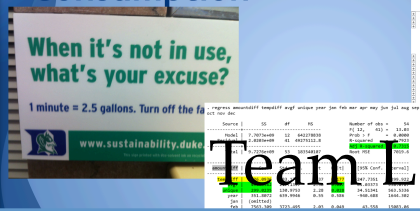
2/2014:
Juarez's Joaquin "El Chapo" Guzman is arrested

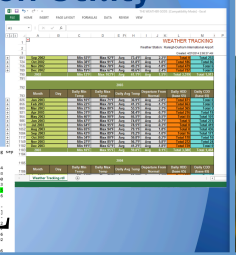
The Middlemen

How does US foreign aid affect underdeveloped countries?

CompSci 290.01 Spring 2014

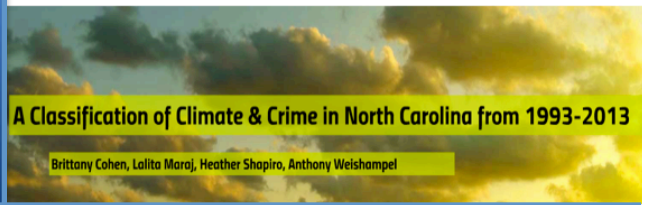
Effects of Renovation on Utility Consumption





Team L

Climate Change



Misc. course info

- **Website**: http://sites.duke.edu/compsci216_01_s2015/
  - Schedule (with links to lecture slides, labs, homework, and additional readings)
  - Help (office hours and online docs)
- **Grading**
  - Project: 50%
  - Homework: 35%, each graded on an X/I/V/E scale
  - Class participation: 15%
    - We'll take lab attendance!
- **Sakai** for grades
- **Piazza** for discussion

Duke Community Standard

- See course website
- Group discussion for homework/labs is okay (and encouraged), but
  - Acknowledge help you receive from others
  - Make sure you "own" your solution
- All suspected cases of violation will be aggressively pursued

More on this topic next Monday

Announcements (Wed. Jan. 7)

- **Homework #1** due next Tuesday midnight
  - See website for details (to be posted by tomorrow night)
  - Short self-intro (submission required)
  - Set up course VM (virtual machine), and play with **OpenRefine** and **regular expressions** for data wrangling