Introduction

Data Cleaning & Integration

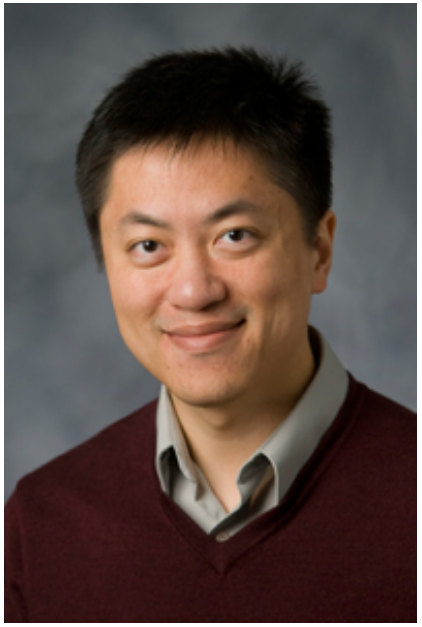
CompSci 590.01 Spring 2017

DUKE
COMPUTER SCIENCE

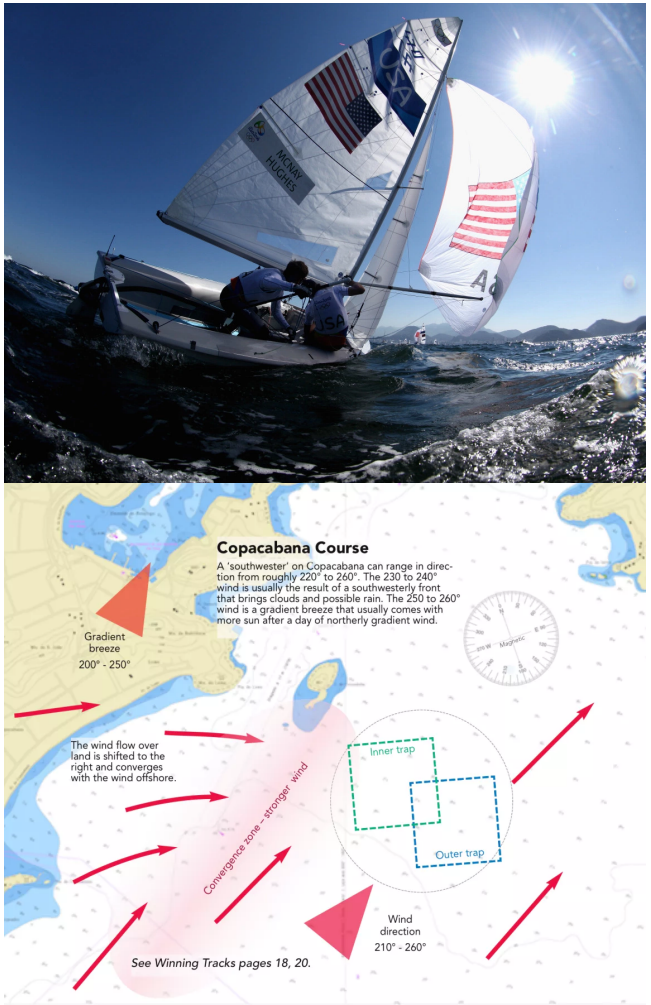
Some contents were based on:
John Canny's CS 194 Fall 2014 slides at UC Berkeley
Xu Chu et al.'s data cleaning tutorial at SIGMOD 2016
Getoor & Machanavajjhala's entity resolution tutorial at VLDB 2012

About me

- Instructor: Jun Yang
  - Been doing (and enjoying) research in databases ever since grad school (1995)
  - Now working on data-intensive systems and computational journalism



Data → gold



… The three years of gathering and analyzing data culminated in what U.S. Sailing calls their "Rio Weather Playbook," a body of critical information about each of the seven courses only available to the U.S. team…

— FiveThirtyEight, "Will Data Help U.S. Sailing Get Back On The Olympic Podium?"

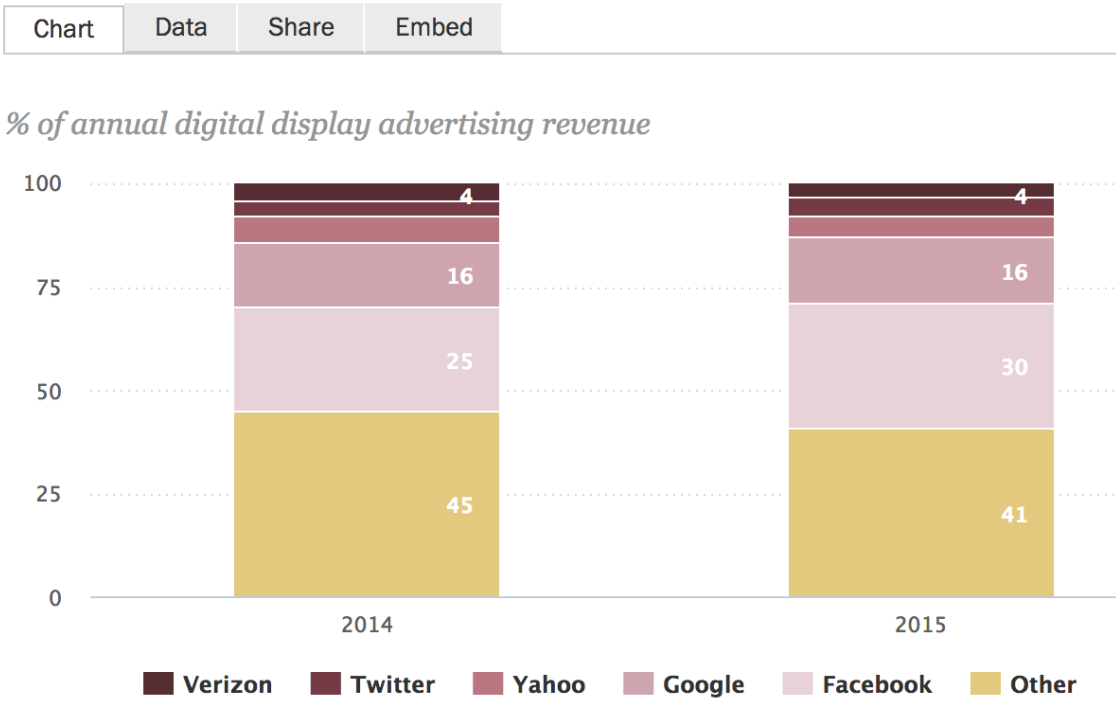
http://fivethirtyeight.com/features/will-data-help-u-s-sailing-get-back-on-the-olympic-podium/

Data → profit

In 2015, $59.6 billion was spent on any digital advertising, including on search engines, social media, news or any other kind of website. This is up 20% from 2014, according to estimates by eMarketer.

— Pew Research

Top five companies account for more than half of total display revenue

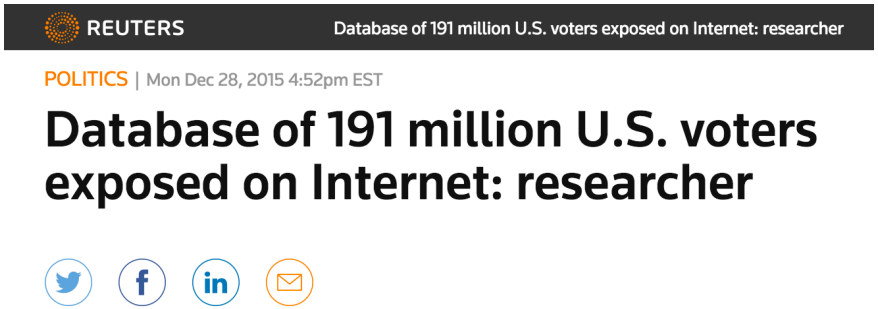


From 2014 to 2015 Google's annual revenue grew over 13%. During that time, revenue from Google websites has comprised a relatively consistent 67 to 68 percent of total company revenue. With the inclusion of the advertising network, Google earns about 90 percent of its entire income from advertising.

— Investopedia

http://www.journalism.org/2016/06/15/digital-news-revenue-fact-sheet/
http://www.investopedia.com/articles/investing/020515/business-google.asp

Data → power

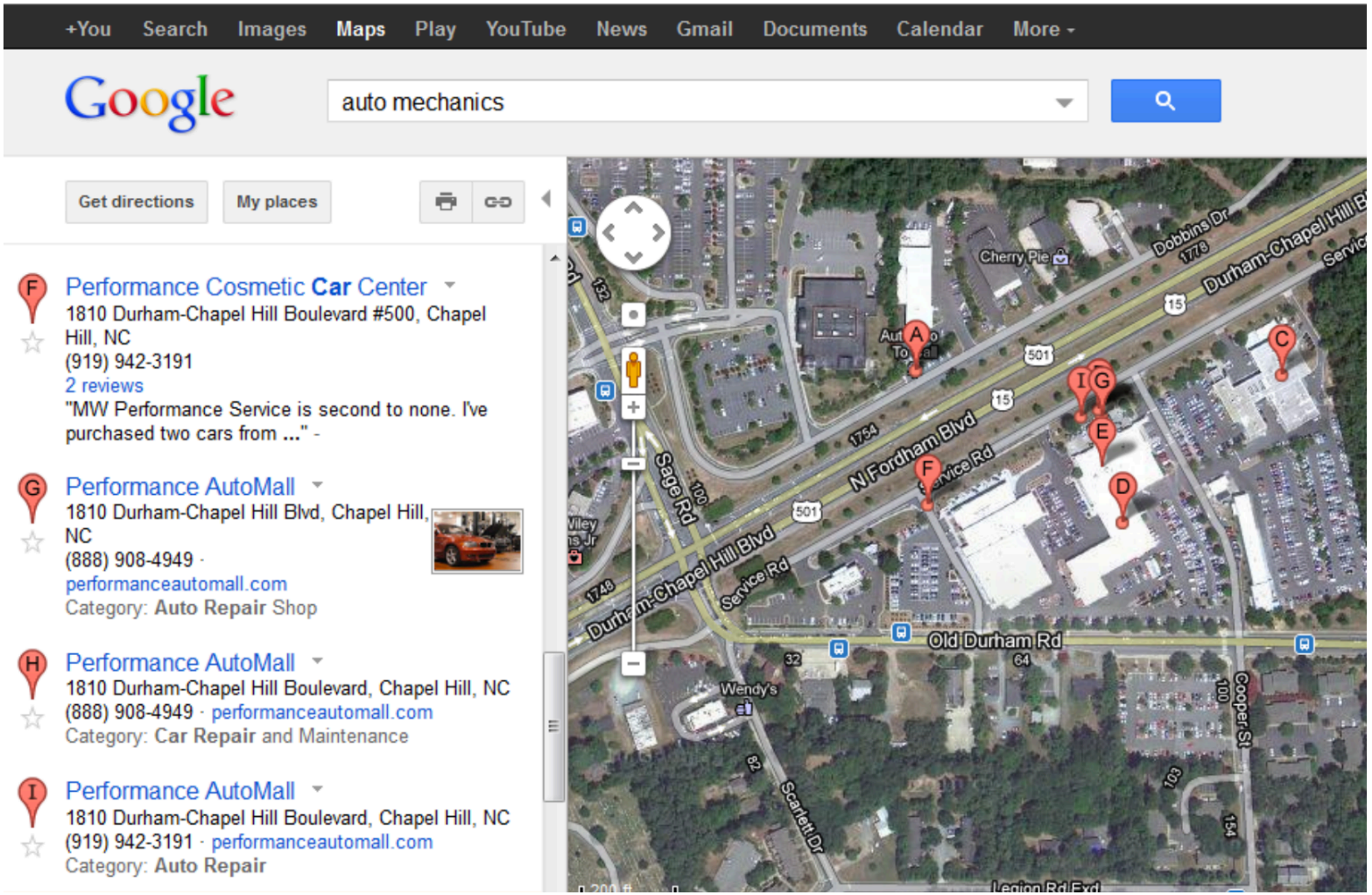


… A database with information on all American voters… might go for about $270,000, according to one marketing firm consulted by researcher Chris Vickery…

— databreaches.net

http://www.npr.org/sections/alltechconsidered/2016/06/24/480949383/britains-google-searches-for-what-is-the-eu-spike-after-brexit-vote
http://www.reuters.com/article/us-usa-voters-breach-idUSKBN0UB1E020151228
https://www.databreaches.net/191-million-voters-personal-info-exposed-by-misconfigured-database/

But without data quality…



(Retrieved in 2015 – they've fixed the problem since then!)

Los Angeles Times

Highest crime rate in L.A.? No, just an LAPD map glitch

By Ben Welsh and Doug Smith

APRIL 5, 2009

On Monday it was a grand theft auto and two robberies, on Tuesday two more robberies and four aggravated assaults. By Friday the toll had risen to 39 major crimes.

And, according to the Los Angeles Police Department's website, that week late last month was pretty typical of the mayhem around the corner from City Hall.

Unable to parse the intersection of Paloma Street and Adams Boulevard, for instance, the computer used a default point for Los Angeles, roughly 1st and Spring streets.

Mistakes could have the effect of masking real crime spikes as well as creating false ones.

http://www.latimes.com/local/la-me-geocoding-errors5-2009apr05-story.html

How Data Failed Us in Calling an Election

By STEVE LOHR and NATASHA SINGER NOV. 10, 2016

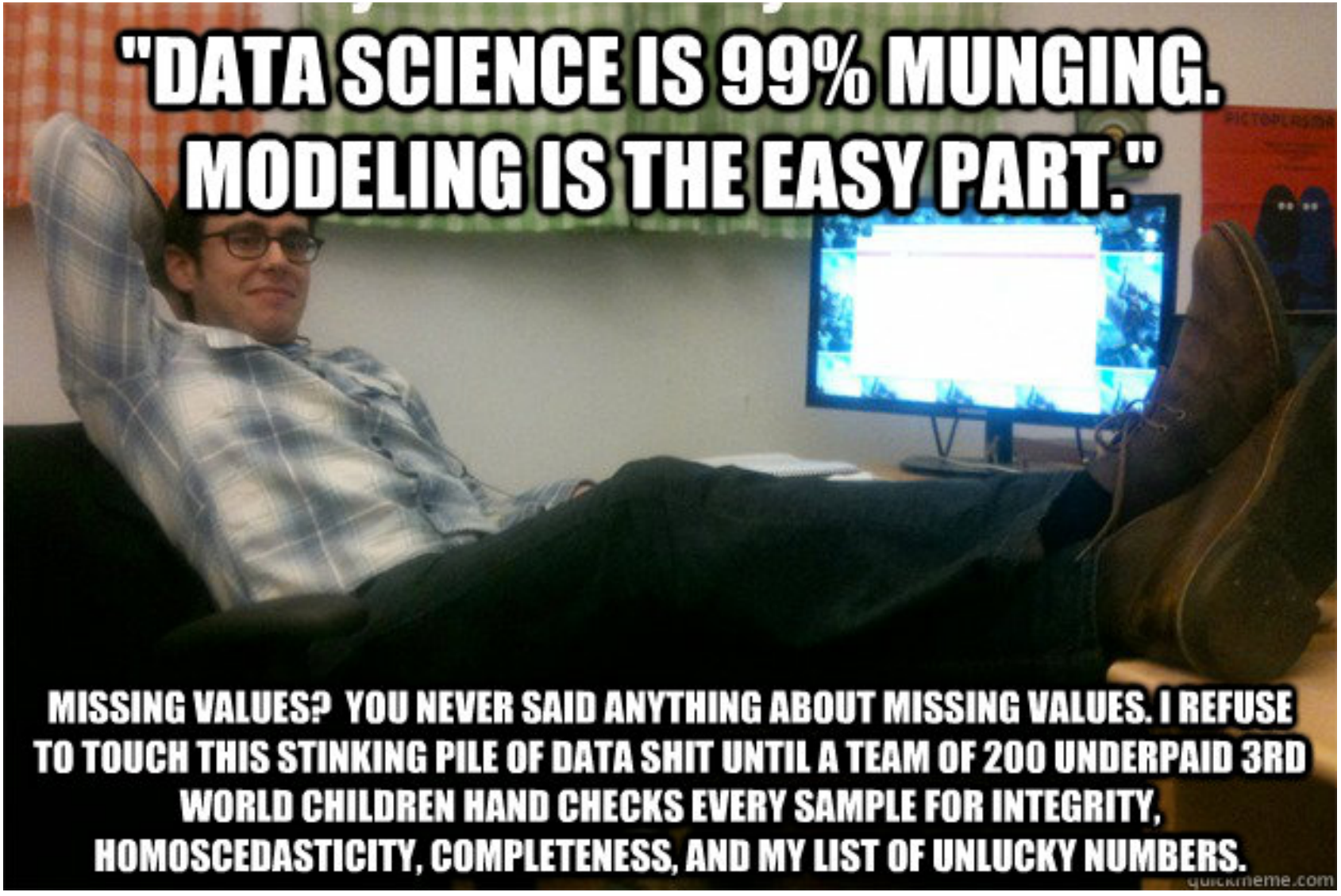
Michigan Democratic Party members in Flint studying precinct results on Tuesday. Virtually all major vote forecasters put Hillary Clinton's chances of winning in the range of 70 to 99 percent.

Brittany Greeson for The New York Times

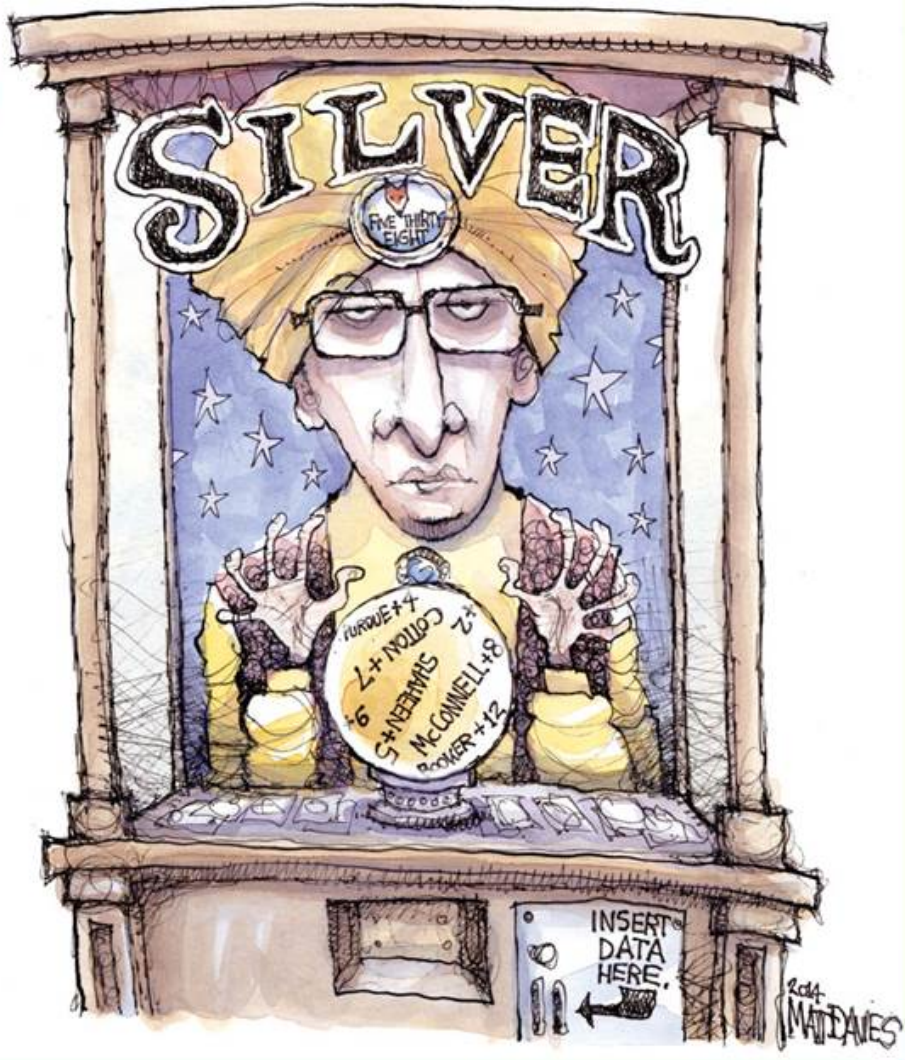
It was a rough night for number crunchers. And for the faith that people in every field — business, politics, sports and academia — have increasingly placed in the power of data.

http://nyti.ms/2eELbOL

Data wrangling/munging



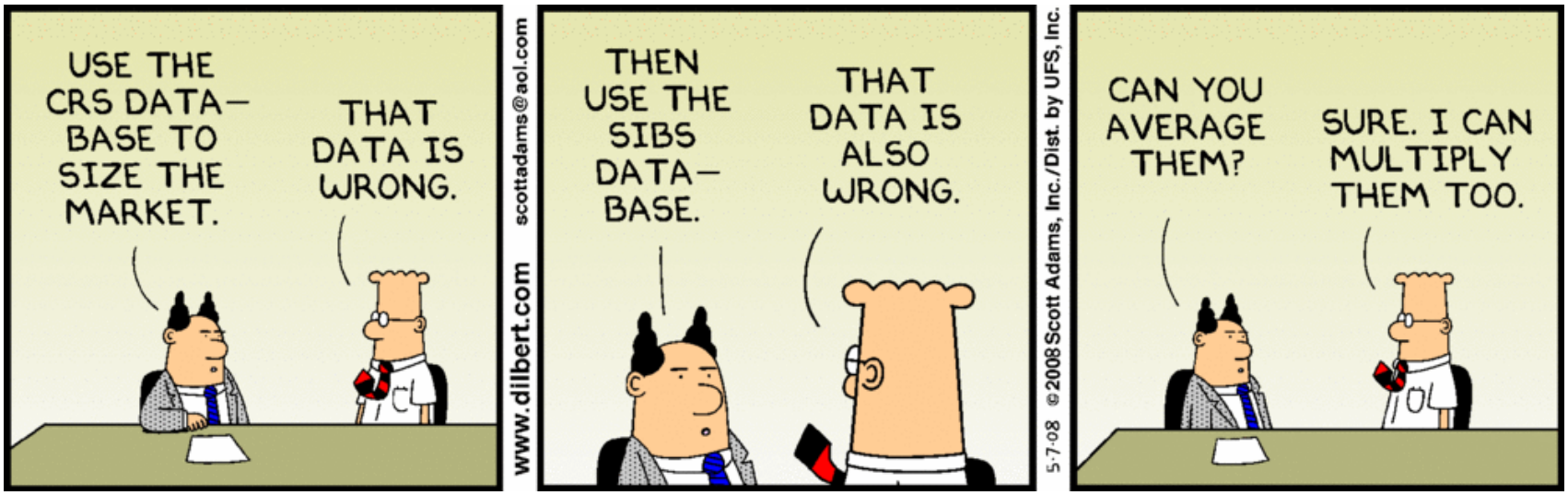
For Nate Silver



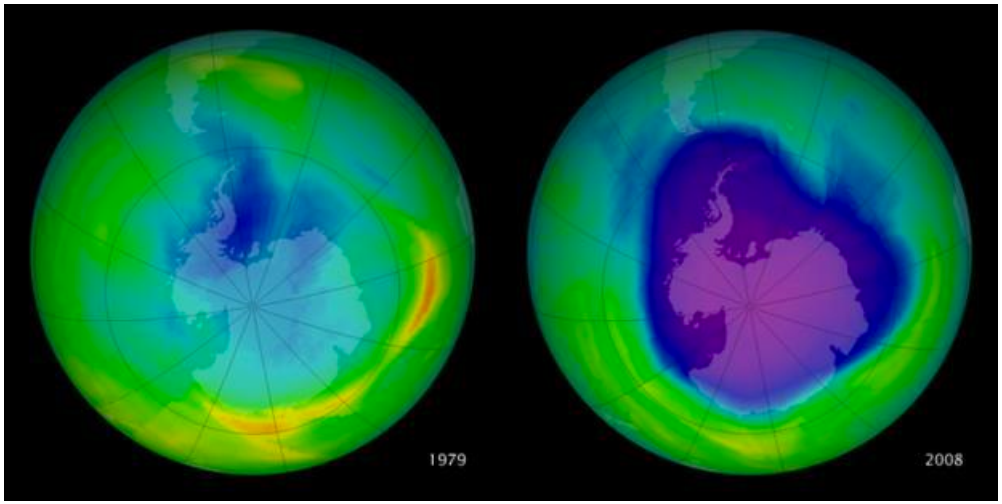
- 70% of the time is spent on getting and cleaning data
- 15% on modeling
- 15% on programming

Personal communication, Nov. 22, 2014

How do we work with dirty data?



Does cleaning makes it okay?



Correct data were rejected as dirty by data quality control algorithms!

The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them; they thought their instruments were malfunctioning.

– National Center for Atmospheric Research

https://commons.wikimedia.org/wiki/File:Agujero_en_la_capa_de_ozono_2008.jpg

Who's calling who's data dirty?

Database view

- You got your hands on this data set (or data sets)
- Some of the values are missing, wrong, inconsistent, duplicated…
- Results to questions (database queries) are absolute
- You can hope to get a “better” answer by improving the quality of values in your data
- Goal: make data hold the truth

Domain experts' view

- You take a look, and …
  - This data doesn't look "right"
  - This answer doesn't look "right"
  - What happened?
- Domain experts have an implicit model of the data that they can test against
- Example: a dissertation proposal began by citing "*Every year since 1950, the number of American children gunned down has doubled*"

– *Damned Lies and Statistics*, by Joel Best

(Example is a more of a miscitation problem, but does illustrate how domain knowledge helps.)

Statisticians' view

- There is a process that produced the data
- In practice samples of that process are non-ideal
  - Distortion: some samples are corrupted by some process
  - Selection bias: likelihood of a sample depends on value
  - Left and right censorship: limited measurement range, or individuals come and go from our scrutiny
  - Dependence: samples are supposed to be independent, but are not (e.g. social networks)
- Impossible to model every type of imperfection
- Goal: find the right trade-off between simplicity and accuracy

Common data quality problems

- "Source" data is dirty on its own
- Transformations corrupt the data (complexities of software pipelines)
- Data sets are clean by themselves but integration screws them up
  - Because of heterogeneity
- "Rare" errors can become frequent after transformation or integration
- Data sets are clean but suffer "bit rot"
  - Old data lose its value over time

Challenges

- How to capture/use prior/domain knowledge
  - Both data and metadata
  - Common sense, business logic
  - Hard vs. soft constraints
- How to work with imperfect data?
  - Clean upfront, or live with imperfection?
  - Focus on what you need in the end
  - Adapt query/analysis to work with uncertainty
- How to scale up?
- How to engage experts?

Teaser problem #1

	ID	FN	LN	Role	Zip	State	Salary
t_1	105	Anne	Nash	E	85376	NY	110
t_2	211	Mark	White	M	90012	NY	80
t_3	386	Mark	Lee	E	85376	AZ	75

- Suppose we know (or really?)
  - No zip code spans multiple states
  - Within each state, a manager (Role='M') should earn no less than an employee (Role='E')
- Is there any hope to automatically suggest "fixes"?
- Are there multiple correct fixes?
- What are good ways to involve users in cleaning (or data entry in the first place)?

Teaser problem #2

Name	twitter
Jacky Spier	@RepSpeier
Tom Price	@RepTomPrice
David Price	@RepDavidEPrice
…	…

bioguide	FN	LN	…
P000523	David	Price	…
P000591	Tom	Price	…
S001175	Jackie	Speier	…
…	…	…	…

- How would you "join" the tables to associate twitter ids with the correct politicians?
  - Exact string comparison?
  - Edit distance?
  - n-gram + Jaccard distance ($|A \cap B|/|A \cup B|$)?
- How fast can you do the matching?
- What if entities to be matched have other features?

A holistic approach

- Consider all steps of the data analysis pipeline
  - How are errors introduced?
  - What results are needed?
  - What can be done at each step of the pipeline?
  - Can we iterate?
- Bring together different perspectives/techniques
  - Database/algorithms
  - Stats/machine learning
  - HCI (Human-Computer Interaction), visualization, crowdsourcing
- Sample papers from different areas

Course load

- Class attendance (10%)
- Reading assignments (25%) throughout the semester, some requiring reviews (submit on Piazza)
- Paper/topic presentations (25%): each student will be expected to present and/or lead discussion in 1-2 class meetings
- Project (40%): teamwork (up to 3 members/team) throughout the semester
- No exams

Grading

[90%, 100%]	A- / A / A+
[80%, 90%)	B- / B / B+
[70%, 80%)	C- / C / C+
[60%, 70%)	D
[0%, 60%)	F

- No “curves”

Misc. course info

- Website: http://sites.duke.edu/compsci590_01_s2017/
  - Office hours; course info; schedule and reading list; lecture slides; …
- Q&A and paper reviews on Piazza
  - See website for link
- Also, please watch your @duke.edu address for announcements

Duke Community Standard

- See course website for link
- Referring to existing literature and group discussion for assignments/project are okay and encouraged, but
  - Acknowledge any help you receive from others
  - Make sure you "own" your solution
- All suspected cases of violation will be aggressively pursued