Introduction

Everything Data
CompSci 290.01 Spring 2014
About us

• Instructors
  – Ashwin Machanavajjhala: data privacy, massive data analytics
  – Jun Yang: data-intensive systems, computational journalism

• TAs
  – Austin Alexander: statistical methods in population statistics
  – Brett Walenz: computational journalism
Let’s talk about $$$

Google to Rake in 33% of Online Ad Revenues This Year
Worldwide digital advertising revenue forecast for the 10 largest ad publishers in 2013

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

**Detecting influenza epidemics using search engine query data**

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

**Red**: official numbers from Center for Disease Control and Prevention; weekly

**Black**: based on Google search logs; daily (potentially instantaneously)
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.whitehouse.gov/blog/Democratizing-Data

Data and culture

• Word frequencies in English-language books in Google’s database

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...

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

"DATA SCIENCE IS 99% MUNGING. MODELING IS THE EASY PART."

MISSING VALUES? YOU NEVER SAID ANYTHING ABOUT MISSING VALUES. I REFUSE TO TOUCH THIS STINKING PILE OF DATA SHIT UNTIL A TEAM OF 200 UNDERPAID 3RD WORLD CHILDREN HAND CHECKS EVERY SAMPLE FOR INTEGRITY, HOMOSCEDASTICITY, COMPLETENESS, AND MY LIST OF UNLUCKY NUMBERS.
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.
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
What happens...

... if you ignore some of these skills
Why this course?

• No single course at Duke gave you the overall picture—we decided 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 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
Misc. course info

• **Website**: [http://sites.duke.edu/compsci290_01_s2014/](http://sites.duke.edu/compsci290_01_s2014/)
  – 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%

• **Sakai** for grades

• **Piazza** for discussion
Duke Community Standard

• See course website for link
• 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
Announcements (Thu. Jan. 9)

• **Homework #1** due Monday midnight: see website for details
  – Short self-intro (submission required)
  – Learn **OpenRefine** and **regular expressions**
    • No submission required for this part unless you want to earn an “E”

• **Top-10** on the wait list should have a good chance of getting in the class
  – Permission #’s will be given next Tuesday
A quick survey

Please raise your hand if you know this term:

- for loops
- Tag clouds
- Inverted lists
- Mashup
- API
- GROUP BY
- Standard deviation
- Bayes’ Rule
- \(k\)-means
- MapReduce