Introduction

Everything Data

CompSci 216 Spring 2019
About us: Instructors

Ashwin Machanavajjhala: big-data, small devices and individual privacy

Jeff Forbes: Computer science education, social information processing, learning analytics
About us: Grad TAs

David Pujol: fairness in differential privacy

Prajakta Kalmegh: cluster dataflow perf

Tiangang (Tim) Chen: ML & distributed systems
About us: Grad TAs

Chenghong Wang: privacy, applied crypto, and biomedical informatics
About us: UTAs

Tatiana Tian: Computer & data science

Muhammad Murtaza: computational finance and databases

Annie Yin: CompSci / AMES
About you…

- On a card:
  1. Preferred Name
  2. Pronunciation
  3. Preferred Pronoun
  4. NetID
  5. Data Science interest
  6. Something interesting and distinctive about you
Tell me …

… why do you want to take Everything Data?
Data and $$$

TOP 10: GLOBAL ADVERTISING REVENUE (IN BILLIONS)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabet</td>
<td>$43.7</td>
<td>$51.1</td>
<td>$59.6</td>
<td>$67.4</td>
<td>$79.4</td>
</tr>
<tr>
<td>Comcast</td>
<td>$11.5</td>
<td>$10.7</td>
<td>$11.8</td>
<td>$17.1</td>
<td>$26.9</td>
</tr>
<tr>
<td>CBS</td>
<td>$8.5</td>
<td>$8.8</td>
<td>$11.5</td>
<td>$11.5</td>
<td>$12.9</td>
</tr>
<tr>
<td>Disney</td>
<td>$7.8</td>
<td>$8.0</td>
<td>$8.2</td>
<td>$10.3</td>
<td>$10.4</td>
</tr>
<tr>
<td>21st Century</td>
<td>$7.6</td>
<td>$7.6</td>
<td>$8.1</td>
<td>$8.5</td>
<td>$8.6</td>
</tr>
<tr>
<td>iHeartMedia</td>
<td>$6.0</td>
<td>$7.0</td>
<td>$7.8</td>
<td>$7.6</td>
<td>$7.8</td>
</tr>
<tr>
<td>Viacom</td>
<td>$4.8</td>
<td>$6.1</td>
<td>$7.2</td>
<td>$5.1</td>
<td>$7.7</td>
</tr>
<tr>
<td>Bertelsmann</td>
<td>$4.7</td>
<td>$5.1</td>
<td>$6.1</td>
<td>$5.8</td>
<td>$6.3</td>
</tr>
<tr>
<td>TimeWarner</td>
<td>$4.3</td>
<td>$4.9</td>
<td>$5.0</td>
<td>$5.0</td>
<td>$5.1</td>
</tr>
<tr>
<td>Facebook</td>
<td>$4.3</td>
<td>$4.6</td>
<td>$4.6</td>
<td>$4.7</td>
<td>$5.1</td>
</tr>
</tbody>
</table>

SOURCE: Bloomberg, Zenith Media
Data collection

01 Activity History
02 Google AdWords Profile
03 Location History
04 Audio History
05 Device History
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 culture

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

Data and social good

Human Trafficking makes more money than Google, Starbucks, Nike, and the NFL Combined

$150 Billion/Year

$136.6 Billion
Combined 2015 Revenue

[Logos and images related to the content]
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

"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.
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 Spark,
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.
The dark side of the force...
Easy to get it wrong …

• Rhine Paradox of extrasensory perception
  http://lastinggems.wordpress.com/tag/rhine-paradox/

- 10 cards either Red or Blue
- Asked people to guess the color of all cards
- 1 in 1000 guessed all correctly – They are told: “You have ESP!!!”
- People with ESP were administered test again, they got it wrong!
Easy to get it wrong …

- Rhine Paradox of extrasensory perception
  http://lastinggems.wordpress.com/tag/rhine-paradox/

… people lose extrasensory perception if they are told they have it …
Easy to get it wrong …

• How accurate is Google Flu Trends?

http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html
Easy to get it wrong …

• How accurate is Google Flu Trends?

Google’s algorithm was quite vulnerable to overfitting to seasonal terms unrelated to the flu, like “high school basketball.” With millions of search terms being fit to the CDC’s data, there were bound to be searches that were strongly correlated by pure chance, and these terms were unlikely to be driven by actual flu cases or predictive of future trends. Google also did not take into account changes in search behavior over time. After the introduction of GFT, Google introduced its suggested search feature as well as a number of new health-based add-ons to help people more effectively find the information they need. While this is great for those using Google, it also makes some search terms more prevalent, throwing off GFT’s tracking.

https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/
Easy to get it wrong …

Brexit Surprise Due to Analysts' Bias, Not Numbers
http://www.realclearpolitics.com/
Easy to get it wrong ...

Or more famously ...

![Map of the United States with states colored in blue or red]

- Blue states represent one outcome
- Red states represent another outcome
Easy to abuse ... ethics

Facebook emotion study breached ethical guidelines, researchers say

Lack of 'informed consent' means that Facebook experiment on nearly 700,000 news feeds broke rules on tests on human subjects, say scientists

Harvard Researchers Used Secret Cameras to Study Attendance. Was That Unethical?

By Rebecca Koenig and Steve Kolowich | November 06, 2014
Easy to discriminate …

- **Redlining**: the practice of denying, or charging more for, services such as banking, insurance, access to health care, or even supermarkets, or denying jobs to residents in particular, often racially determined, areas.
The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people’s behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

Google’s online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Research from Harvard University found that ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity. The Federal Trade Commission said advertisers are able to target people who live in low-income neighborhoods with high-interest loans.
Recidivism Prediction

Data

Criminal history of defendant (and others)

Decision Maker

Decision

High risk of recommittting a crime. Do not grant bail.

Low risk of recommittting a crime. Grant bail.
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
Easy to invade your privacy …

**A Face Is Exposed for AOL Searcher No. 4417749**

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

**Why 'Anonymous' Data Sometimes Isn't**

By Bruce Schneier
12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

**“Anonymous” Genomes Identified**

The Scientist » The Nutshell

The names and addresses of people participating in the Personal Genome Project can be easily tracked down despite such data being left off their online profiles.

By Dan Cossins | May 3, 2013
US military reviewing security practices after fitness app reveals sensitive info

By Joshua Berlinger and Maegan Vazquez, CNN
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!

Learn to

• Take advantage of it, and
• Help yourself and others avoid being taken advantage of
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
Is this course right for me?

• We assume previous experience with programming (ideally in Python) at the level of 201
  – If not, you might want to take CS116

• We assume knowledge of probability and basic statistics
  – If not, you might want to take STAT courses like 111, 199, 230 or CS230.

• We assume you have not taken advanced CS courses like 316 or 371
  – If you have taken these courses, please meet with the instructors.
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 (Thu) and hands-on, team-based labs (Tue)
  - 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/compsci216s2019/
  – Schedule (with links to lecture slides, labs, homework, and additional readings)
  – Help (office hours and online docs)

• **Sakai** for grades

• **Piazza** for discussion
Grading

• **Individual effort: 50%**
  – 12 Homeworks
    • Based on material covered in class, and setup for lab assignment
    • 10 highest scoring homeworks count for the grade

• **Team effort: 50%**
  – Lab work: 15%
    • Attendance and attempting the lab assignment.

  – Project: 35%
    • Submitting project proposal, midterm and final reports on time
    • Quality of the project work
    • Writing and presentation
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
Announcements (Thu. Jan. 10)

• **Homework #0**
  – Setup course VM
  – Setup git
  – Send us your git credentials

• **Lab #0**
  – *In class “Preparedness” test*
  – Help with VM/git setup