ENSF 612: Engineering Large Scale Data Analytics Systems

Data Analysis and Big Data

Sarah Shah,
Department of Electrical and Software Engineering,
University of Calgary.

Topics

- Course objectives - revisited
- Data analysis - definition
- Big data - definition
- Sources of big data
- Big data case studies

Course objectives - revisited

- What will we learn in this course?
  - What is data analysis (data science)?
  - What is “big” data and big data analysis?
  - What are sources of big data?
  - Why is there so much excitement about it?
  - What are the benefits of analyzing big data?
  - What are some platforms available to develop algorithms to analyze big data?
  - How to write big data programs on these platforms?
  - How to develop data-driven models for prediction?

Course objectives - revisited

- What will we learn in this course?
  - **What is data analysis (data science)?**
  - What is “big” data and big data analysis?
  - What are sources of big data?
  - Why is there so much excitement about it?
  - What are the benefits of analyzing big data?
  - What are some platforms available to develop algorithms to analyze big data?
  - How to write big data programs on these platforms?
  - How to develop data-driven models for prediction?

Data Analysis

- Been around for a long time
- Data science definition:
  - Exploit data to find useful patterns and information
  - Involves methods to
    - **Collect** data
    - **Inspect quality** of data
    - **Extract subsets** of data
    - **Transform** data
    - Do **Exploratory Data Analysis**
    - **Build models** from data

Data Analysis - Example

- Data analysis process with an example Deerfoot Trail traffic analysis
  - City planners interested in following questions
    - What is commute time from point A to B on the highway?
    - How does this commute time change over the day?
    - How is the time impacted by factors weather and accidents?
    - What are the most congested sections of the highway?
    - How do these sections change with time and season?
    - Can we predict future commute times?

What are the various steps in this analysis?

Data Analysis - Example

- Figure out data sources – collect data
  - Commute time – road sensors, Google maps
  - Weather – Environment Canada
  - Accidents – police reports, tweets from users
- Inspect data quality
  - Check for duplicate or missing or incorrect data
  - E.g., no +30 C days in January
  - Does sensor data agree with Google maps?

Data Analysis - Example

- Extract subset of data
  - Which year do we want to study?
  - Which roads do we want to study?
  - Which seasons do we want to study?
- Transform data
  - Merge data – transform it to format good for analysis tools
  - E.g., Excel likes data in comma-separated format
    - Date,Time,Highwaysection,commutetime,#ofaccidents, temperature,snowonground,…….

Data Analysis - Example

- Do preliminary or exploratory data analysis (EDA)
  - Compute statistics such as mean and median
  - Visualize data for weekends, weekdays, winter, etc.
  - EDA can help understand how to answer our questions
  - EDA can help us build **models**
- Build models
  - Predict output variable as a function of input variables or features

    Example
    - Output variable – commute time
    - Features – day, time, accidents, weather
    - Models need to be **trained** and **validated**

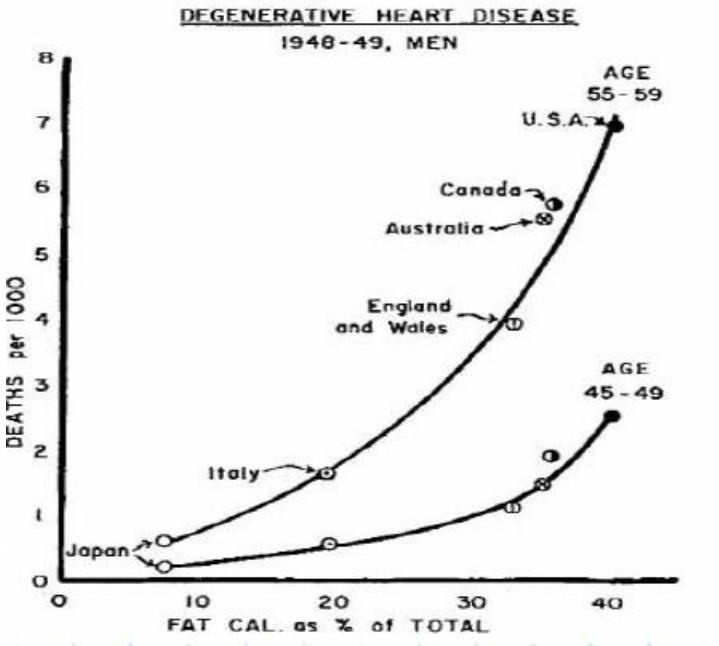
Data analysis process is typically iterative

Data analysis pitfalls

- Trap of falling in to “correlation is causation”
  - Say variables a and b are positively correlated
  - Temptation might be to conclude
    - “as a increases b increases”
    - “as a decreases b decreases”
  - This might not be necessarily true!
  - Let’s look at a few examples

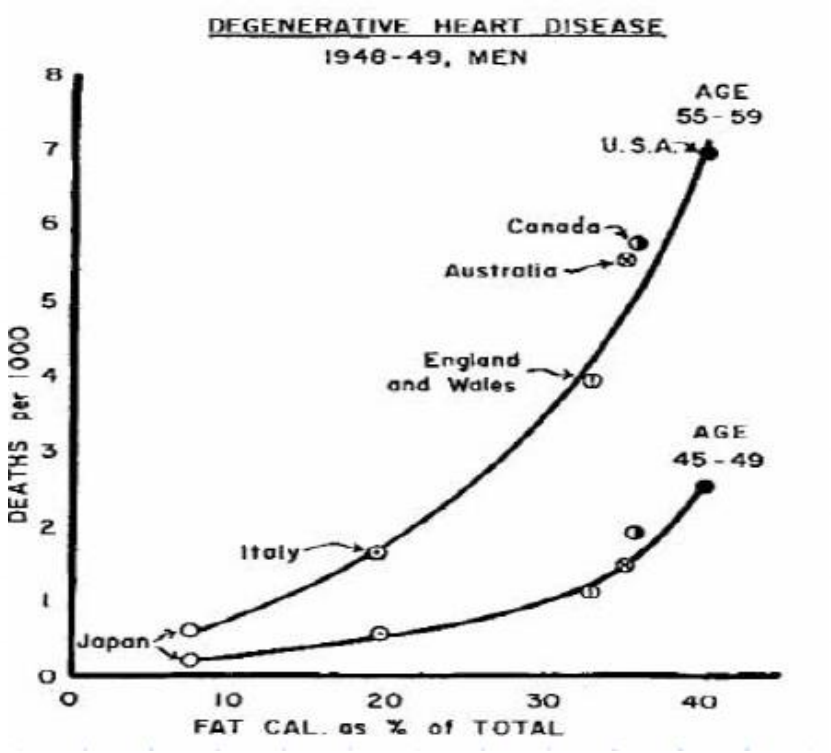
Data analysis pitfalls - examples

- Seven countries Study – Ancel Keys
- Started in 1958 -- followed 13,000 subjects between the ages of 5-40



Data analysis pitfalls – examples

- Significant controversy
- Failed to consider other factors, e.g., sugar consumption
- **Correlation is not causation!**



Data analysis pitfalls - examples

- "Epidemiological modeling of online social network dynamics" by Cannarella and Spechler from Princeton

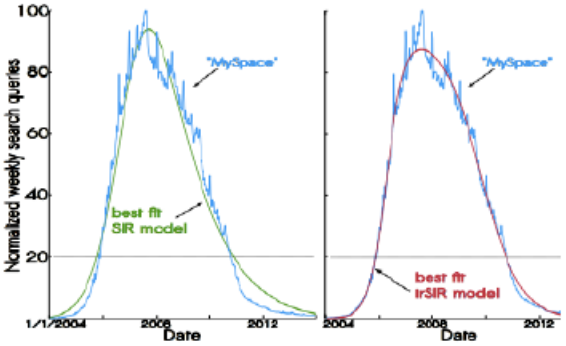
- The following was derived from the abstract:
  - "Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years"

How did they arrive at this conclusion?

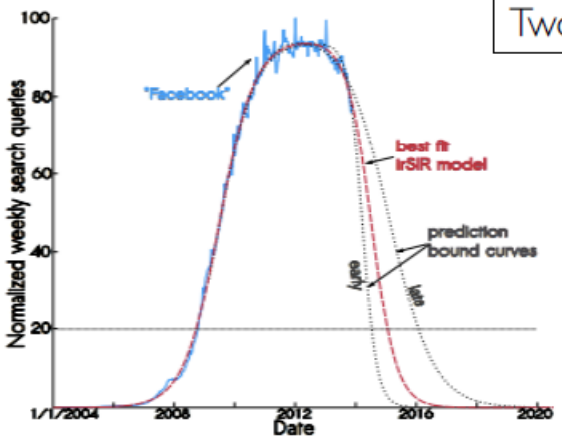
Paper available at http://arxiv.org/abs/1401.4208

Data analysis pitfalls - examples

Google Trends searches for "MySpace"



Two Figures from the paper



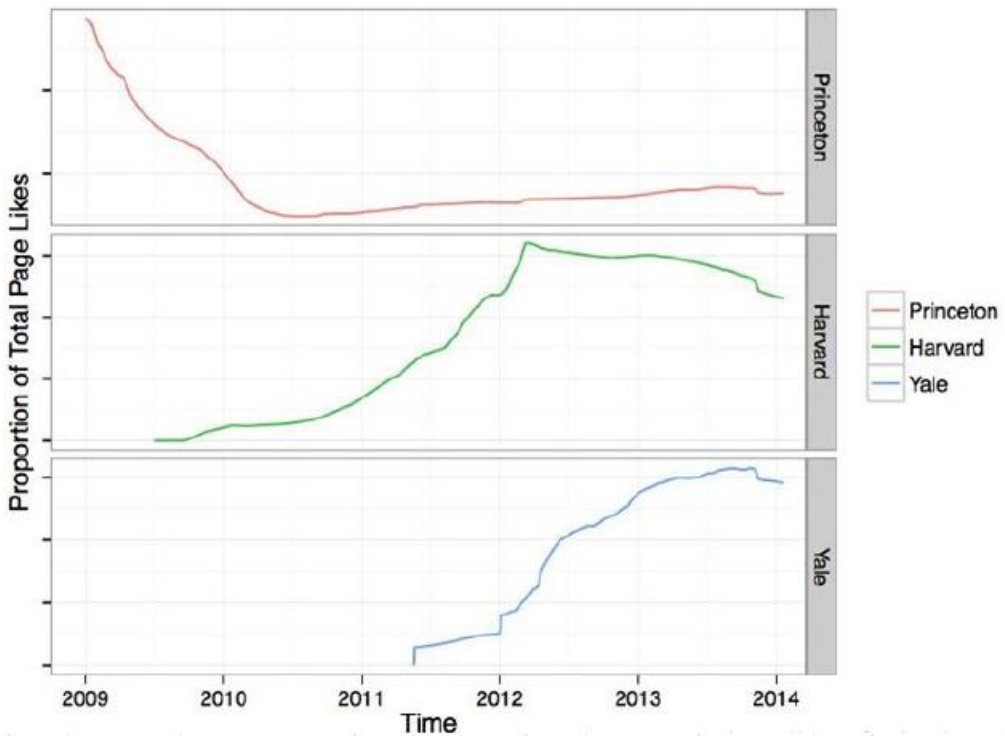
Searches for "Facebook"

- Declining search correlates to declining popularity

- **Falling prey to correlation is causation!**

Data analysis pitfalls - examples

- Facebook's response:



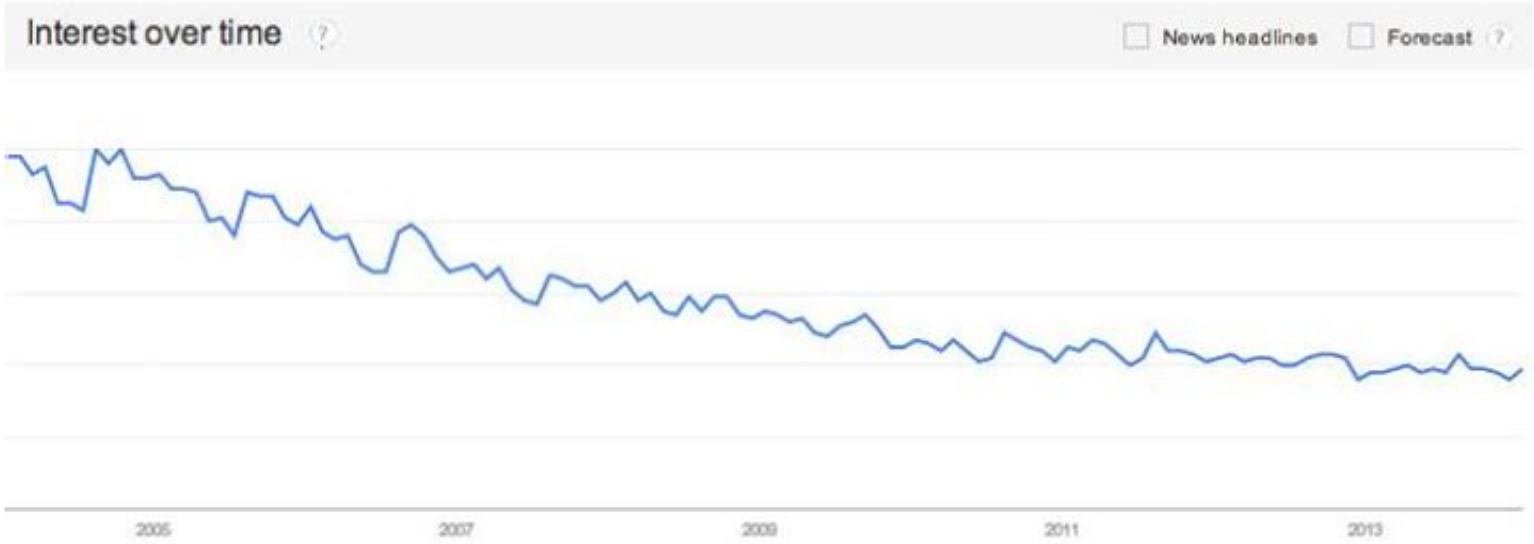
In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely.

https://www.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849/

Data analysis pitfalls - examples

- Facebook's response:

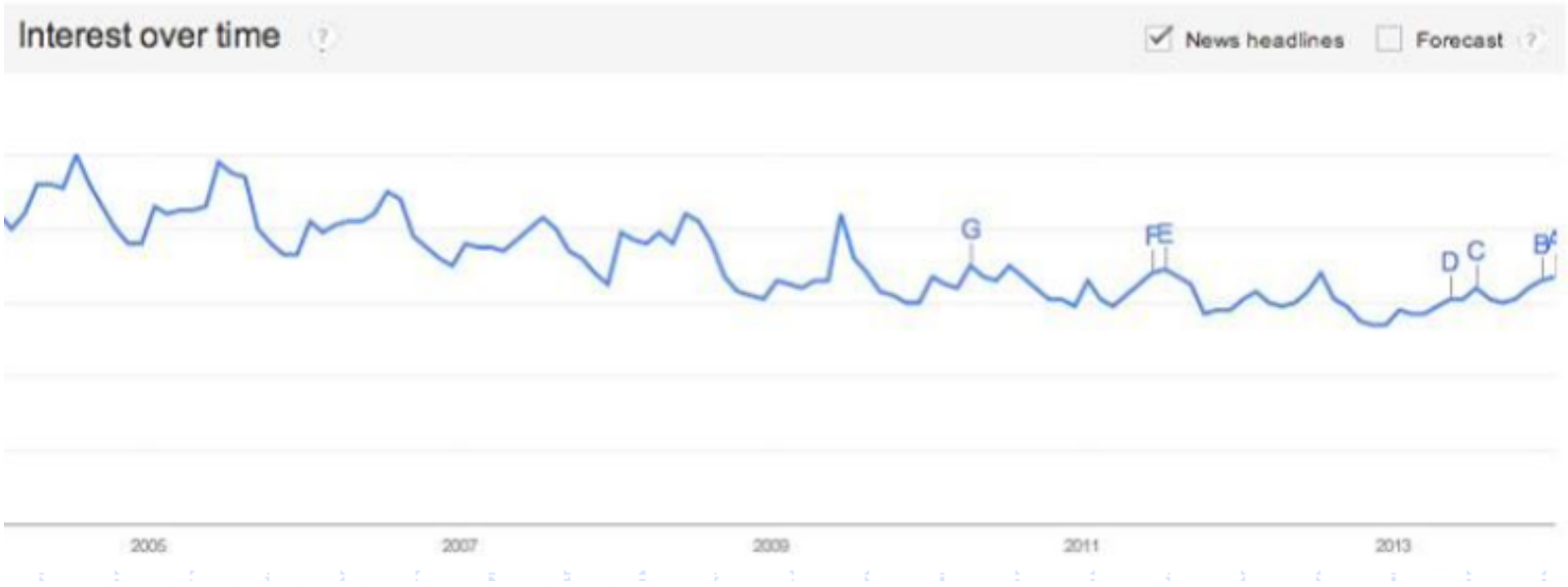
> … and based on Princeton search trends:
>
> "This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…"



Data analysis pitfalls - examples

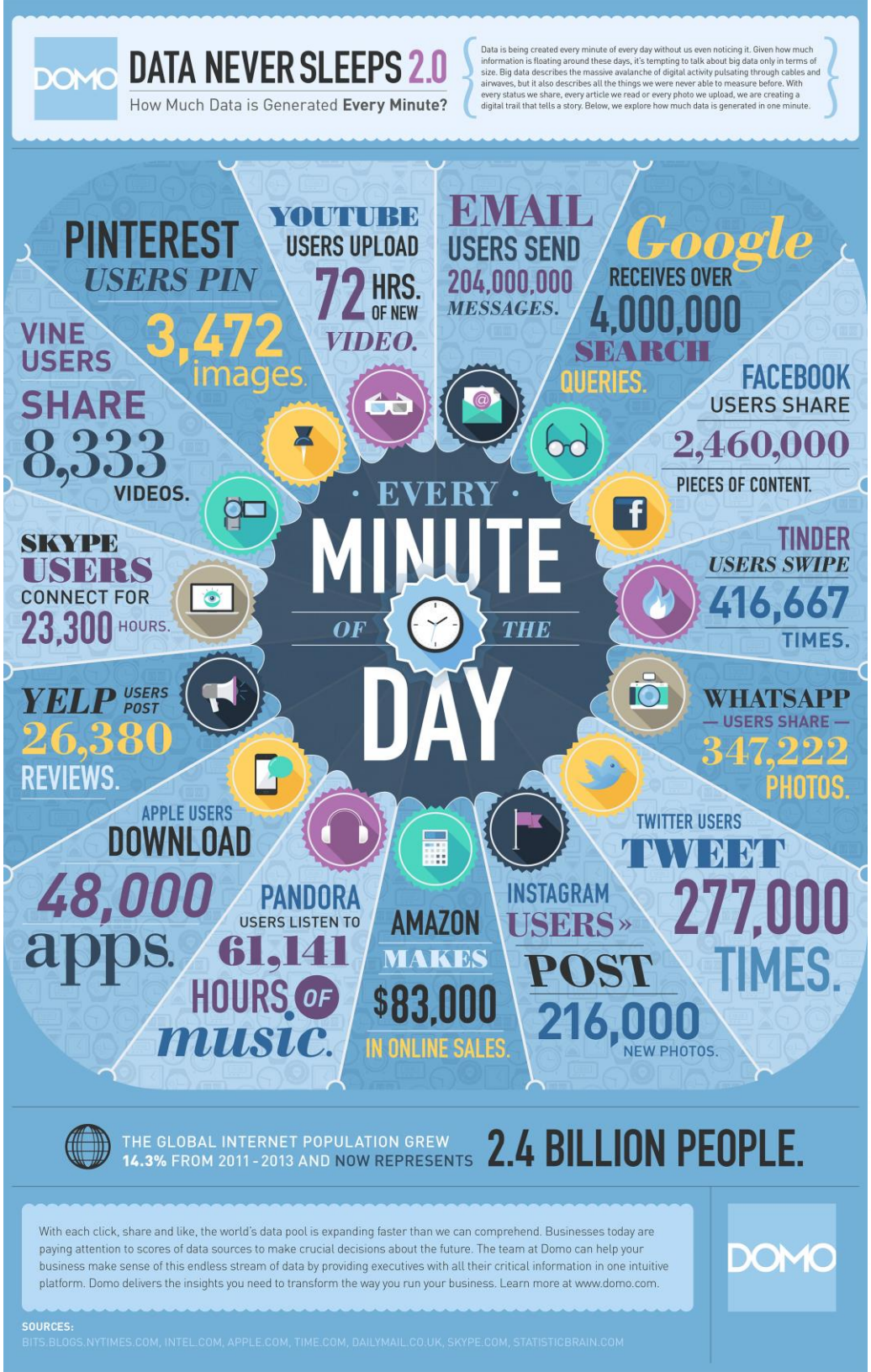
- To further drive the point home, Facebook pointed out:

> While we are concerned for Princeton University, we are even more concerned about the fate of the planet — Google Trends for "air" have also been declining steadily, and our projections show that by the year 2060 there will be no air left:



Course objectives - revisited

- What will we learn in this course?
  - What is data analysis (data science)?
  - **What is "big" data and big data analysis?**
  - What are sources of big data?
  - Why is there so much excitement about it?
  - What are the benefits of analyzing big data?
  - What are some platforms available to develop algorithms to analyze big data?
  - How to write big data programs on these platforms?
  - How to develop data-driven models for prediction?



Big data

- Big data is a relatively new term
- Refers to exponential growth in data
  - Data deluge – 2010 cover story in "The Economist"
  - 150 billion GBs of data in 2005 – 1200 billion GBs in 2010
- Image shows scale of the data

Big data – cont'd

- How big is big data?
  - Social networks
    - Twitter – 316 million users; 500 million tweets/ day
    - Facebook – 968 million users; 55 million status updates/day
  - Search
    - Google – 30 trillion pages indexed – 10^8 gigs of index data
    - Google uses 1 million compute hours to build index
  - Science
    - The square kilometer array – 700 TB/sec data to be persisted
  - More examples will follow later…..

How does one define big data?

Big data – cont'd

- Definition of big data?
  - Many definitions exist
  - Coarse definition for this course
    - **"Any dataset that doesn't fit reasonably in a single computer"**
    - Problematic definition since computer specs keep changing
    - Sample high end computer
      - HP Superdome–16 Xeon processors–3 TB RAM
    - Data will expand to make single computer inadequate

Course objectives - revisited

- What will we learn in this course?
  - What is data analysis (data science)?
  - What is “big” data and big data analysis?
  - **What are sources of big data?**
  - **Why is there so much excitement about it?**
  - **What are the benefits of analyzing big data?**
  - What are some platforms available to develop algorithms to analyze big data?
  - How to write big data programs on these platforms?
  - How to develop data-driven models for prediction?

Sources of big data

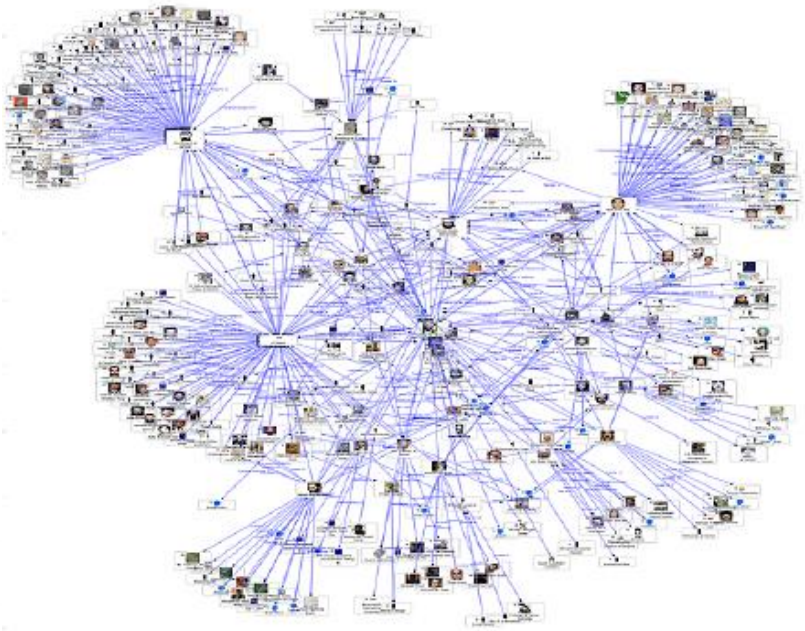
- Actions of Web users
  - Every click, pause/stop/play, ad click stored
  - Data can be analyzed for profit
    - Targeting ads to users (Google)
    - Recommender systems (Netfilix, Amazon)
- Content generated by Web users
  - Tweets, posts, content
  - Individually not much – but together can mean a lot
    - Facebook sells mined info on users to others
    - Graph analytics – identify “communities” in social networks

Sources of big data – cont'd

- Scientific data
  - Square km array
  - Large hadron collider
    - Greater than Wikipedia content, Tweets/day, library of congress
  - Genome sequencing data
    - Increase in genome data due to inexpensive sequencers
    - Human genome typically needs 200GB of storage
    - E.g., large genome databases can be used to detect diseases
- Neuronal data from organisms
  - Identify what neurons are in charge of what functions
  - Larval zebra fish – 100,000 neurons – generates 1 TB/experiment
  - Mouse – 80 million neurons – generates 100 TB/experiment

Sources of big data – cont'd

- Log files from computers in data centers
  - Detection of security and performance problems
- Graph analytics
  - Mining graphs for patterns
    - E.g., friend/account recommendations on social media platforms
    - E.g., commute time prediction in road networks



Big data case studies

- Google's flu trends
  - Traditional methods of predicting outbreak are slow
    - Manually collect reports from clinics/hospitals
  - By the time an alert is issued, outbreak has already happened
  - Google predicted outbreak 2 weeks before government
  - How did they do it?
    - Started with 50 million search queries from 2003-2008
    - Identified 45 terms related to flu used by people with flu
    - Used these terms as features in a model
    - Model predicted flu cases on a week by week basis
    - Model was found to be 97% accurate

https://www.youtube.com/watch?v=6111nS66Dpk

Big data case studies

- President Obama's re-election campaign
  - Big data analytics credited for successful campaign
  - Data models for targeting ads and fundraising
  - Algorithms to identify electors on the fence
- Interview with the data science team

http://poy.time.com/2012/12/19/obamas-data-team/

Big data case studies

- Mining Twitter for customer service
  - Restaurant used big data analysis to satisfy a customer
  - More on the story in these links

http://searchcio.techtarget.com/opinion/Ten-big-data-case-studies-in-a-nutshell

http://shankman.com/the-best-customer-service-story-ever-told-starring-mortons-steakhouse/

Topics

- Course objectives - revisited
- Data analysis - definition
- Big data - definition
- Sources of big data
- Big data case studies

Acknowledgements

Portions of these slides were adapted from external material available under creative commons license CC-BY-NC-SA 4.0. This license grants the ability to share and adapt the material for non-commercial purposes.

Name of the creator: Dr. Anthony Joseph, University of Berkeley and team

License notice: https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

Copyright notice: CC-BY-NC-SA 4.0

Link to material: https://courses.edx.org/courses/BerkeleyX/CS100.1x/1T2015/info