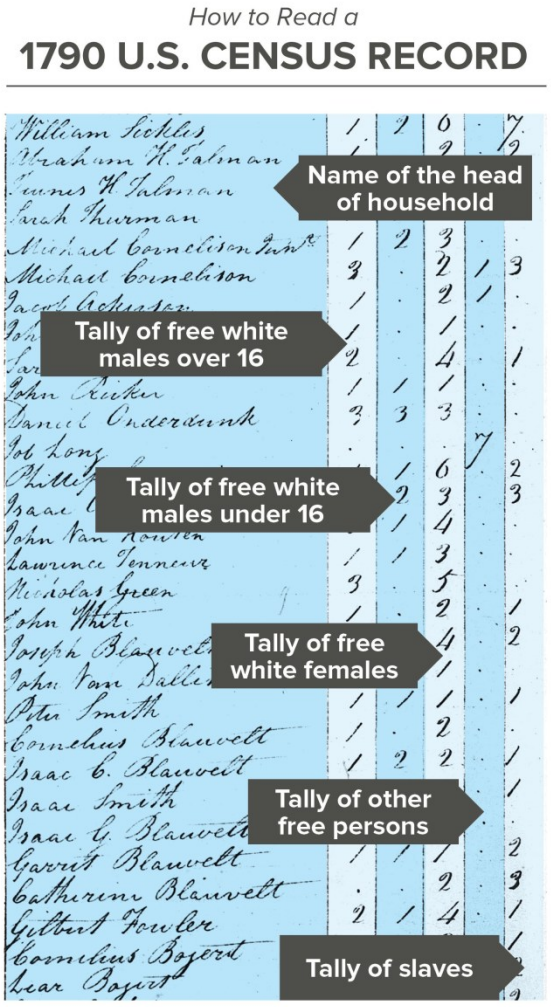
SOSC 4300/5500: Overview

Han Zhang

Outline

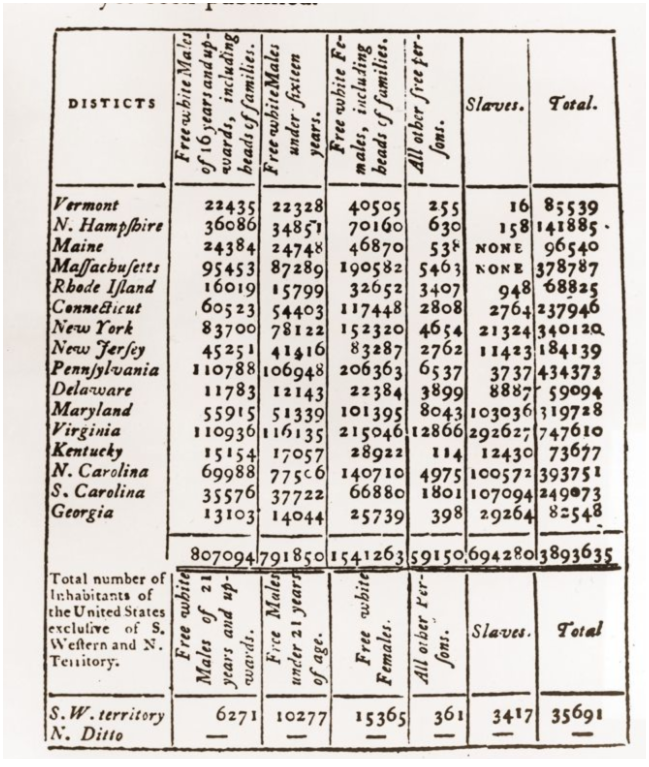
Before Digital Revolution

- https://www.familysearch.org/blog/en/1790-census-form-questions/



Before Digital Revolution

- And then we calculate some statistics from census surveys
- 1890 US census took 8 years to clean and process by humans



DISTICTS	Free white Males of 16 years and upwards, including heads of families.	Free white Males under sixteen years.	Free white Females, including heads of families.	All other free persons.	Slaves.	Total.
Vermont	22435	22328	40505	255	16	85539
N. Hampshire	36086	34851	70160	630	158	141885
Maine	24384	24748	46870	538	NONE	96540
Massachusetts	95453	87289	190582	5463	NONE	378787
Rhode Island	16019	15799	32652	3407	948	68825
Connecticut	60523	54403	117448	2808	2764	237946
New York	83700	78122	152320	4654	21324	340120
New Jersey	45251	41416	83287	2762	11423	184139
Pennsylvania	110788	106948	206363	6537	3737	434373
Delaware	11783	12143	22384	3899	8887	59094
Maryland	55915	51339	101395	8043	103036	319728
Virginia	110936	116135	215046	12866	292627	747610
Kentucky	15154	17057	28922	114	12430	73677
N. Carolina	69988	77506	140710	4975	100572	393751
S. Carolina	35576	37722	66880	1801	107094	249073
Georgia	13103	14044	25739	398	29264	82548
	807094	791850	1541263	59150	694280	3893635

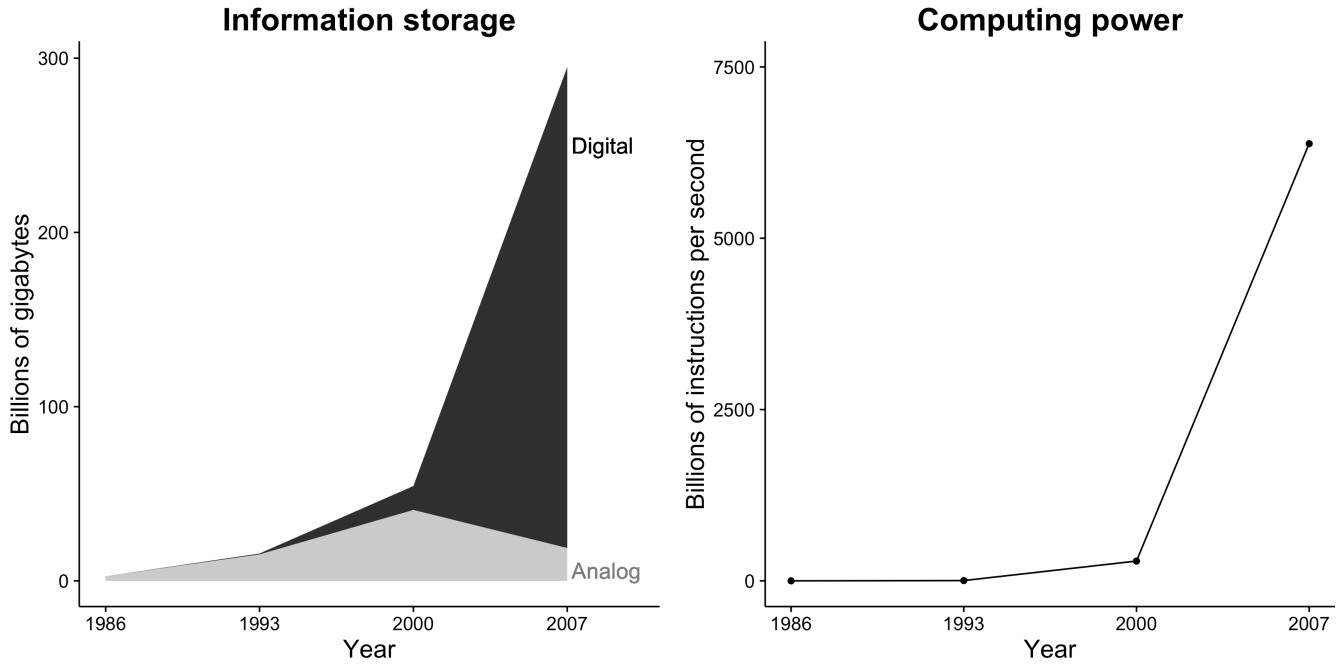
Total number of Inhabitants of the United States exclusive of S. Western and N. Territory.	Free white Males of 21 years and upwards.	Free Males under 21 years of age.	Free white Females.	All other Persons.	Slaves.	Total
S. W. territory	6271	10277	15365	361	3417	35691
N. Ditto	—	—	—	—	—	—

With modern computers

- Invented in 1940s, modern personal computers become popular since 1980s
- Calculation becomes much faster with computers
  - Imagine solving a regression by hand without computers!
- But data are still in analog format; they are represented in a physical way.

Welcome to the digital age

- Since 2000, both computing power and digital data are quickly increasing
- Hilbert, Martin, and Priscila López. 2011. The Worlds Technological Capacity to Store, Communicate, and Compute Information. *Science* 332 (6025):6065.



Welcome to the digital age

- Computers everywhere, digital traces everywhere
  - personal computers, mobile phones, cars, watches, thermostats, CCTV cameras...
  - these devices not only calculate; they also measure and store lots of digital data
  - E.g., 20 years ago, you may walk into a bookstore and browse books; no traces will be left once you walk outside the book store
  - Now, your entire online browsing and purchasing behaviors are stored, and will be used for advertising or recommendation for similar products
  - Digital traces do not need to be on Internet!
    - E.g., octopus card swipes allow companies to locate your moving trajectories
- Digital traces are byproducts of peoples everyday actions, often collected by companies.
  - Before digital age, they just fade away, but now they are kept

Welcome to the digital age

- More and more governments and organizations are also turning traditional analog data into digital data
- From printed newspapers to electronic newspaper databases
- From printed maps to Google maps
- [in class activity]: can you think of other examples of digital data that are transformed from traditional analog data?

Big Data

- Together, we call digital traces and traditional data that are turned into digital data as Big Data
- "Big data are created and collected by companies and governments for purposes other than research"
- Matthew Salganik, *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2019
- `https://www.bitbybitbook.com/en/1st-ed/observing-behavior/data/`

Big data vs traditional social science data

- Traditional social science data are made for research
  - Although the data are small, they are ready to use for examining social science theories
- Big data are repurposed for research
  - They are big
  - But you need some effort to get what you want

Big data need different methods

- Previously social scientists have survey and sometimes small administrative data
  - Using various regression models to analyze the data
- Big data are not only big, but also qualitatively different in their formats:
  - Texts
  - Images, Video, Audio
  - Networks
- Analyzing the above need other methods, in particular machine learning
  - Regression in most times won't work!

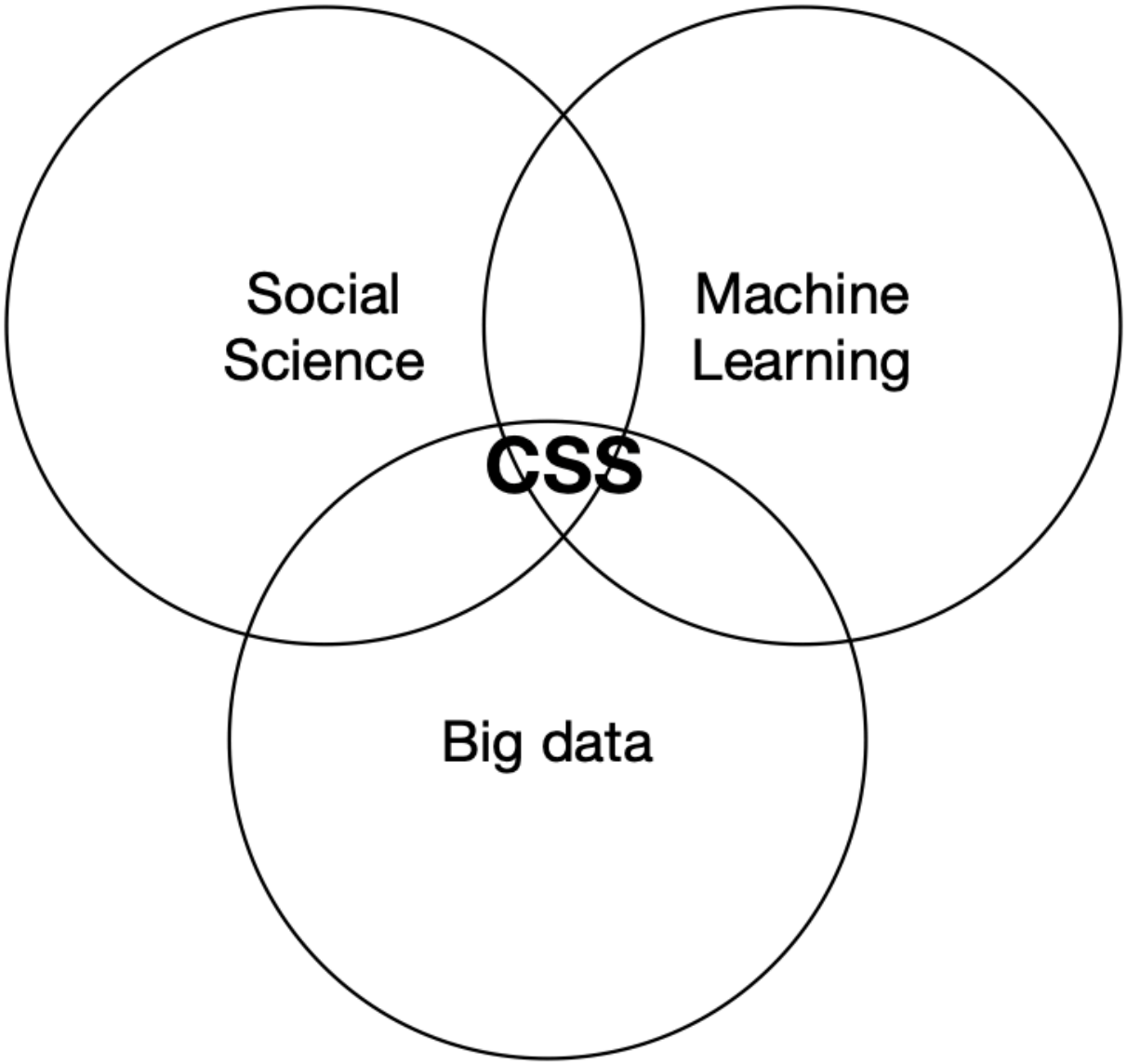
Social scientists and data scientists

- Status quo:
- Social scientists: computational social science
  - Try to turn big data into small data, and then apply traditional regression models
- Data scientists: computational social science
  - Get more data, and apply some fancy machine learning algorithms over social data

Computational Social Science

- Social science itself is not enough, because data can only gets bigger
- Data science itself is not enough, if we want to study social behaviors rigorously
- Computational social science (CSS): bridging social and data science

Computational Social Science (CSS)



Study Goals

1. Describe the opportunities and challenges of computational social science
2. Evaluate computational social science research on social phenomena
3. Practice the essential techniques to analyze social big data, especially text data (covered in Tutorials)
    - Getting data
    - Managing data
    - Analyzing data with appropriate methods (ML, GPT)
4. Propose research questions that are suited to be examined by computational methods with big data
5. Write a research article that utilizes the techniques and methods of computational social science to address social science problems, or design a project that use computational social science to address some real-world problems.

Instructors

Instructor: ZHANG, Han

- Office: 2379
- Email: zhangh@ust.hk
- Office Hour: Mon 3-4PM (or email me to find a time)

Teaching Assistant: Li, Jingchen

- Office: 3001
- Email: jlieg@connect.ust.hk
- Office Hour: TBD

Course material

- Lecture material and tutorial will be available at: `https://github.com/HKUST-SOSC4300-5500`
- Please use the version on GitHub as the authoritative version

Schedule (tentative)

Week	Date	Topic
1	*[2024-02-06 Tue]*	Introduction; big data
2	*[2024-02-20 Tue]*	Prediction;
3	*[2024-02-27 Tue]*	Prediction; Evaluation
4	*[2024-03-05 Tue]*	Text (I); representation
5	*[2024-03-12 Tue]*	Text (II); supervised
6	*[2024-03-19 Tue]*	Text (III); embedding
7	*[2024-03-26 Tue]*	Text (IV); unsupervised
8	*[2024-04-09 Tue]*	Network; basics
9	*[2024-04-16 Tue]*	Network; small worlds
10	*[2024-04-23 Tue]*	Causal Inference and Big Data: network as
11	*[2024-04-30 Tue]*	Image data (or other elective topics)
12	*[2024-05-07 Tue]*	Presentation

Grading Components

	%	Due
Attendence and participation	10%	
Homework assignments	30%	Two weeks
Literature review		
Report	5%	
Presentation	5% (5 min)	mid April
Final Paper/Project		
Presentation	15% (15 min)	May 7
Write-up	35%	May 28

- 4300 and 5500 will be graded independently

Attendance and participation

- Please read the required weekly reading from the syllabus
- I will assume you have read the content and skip these quickly
- I will also ask you questions related to the reading, which will be counted as participation scores

Grouping

- You should finish all tasks in groups
  - If there is any MPhil or PhD student in a group: max group size is 2
  - Otherwise: 3 to 4 in a group (e.g., 4 UG in a group)
- Finish grouping by the end of February
  - we will have first assignment then

Class participation

- Read the required readings on syllabus before each week's lecture
- Answer questions about the assigned readings
- Ask questions about the parts you did not understand.
- If you are uncomfortable speaking up in class, send the question in Zoom's chat window, post them on Github, come to my office hours, or send your questions to instructors via e-mail.

Homework assignments

- There will be 3 to 4 coding exercise as homework.
- These homework assignments test your knowledge of analyzing data using statistical software.
- Each exercise is due in two weeks after the release of assignment.

Literature Review

- Select a social topic and summarize how researchers have used computational methods and/or big data to study this particular research area.
- Be specific: don't choose topics such as "social media texts and sentiment analysis"
- Some examples of research areas:
  - Sociology: internal or international migration, social inequality, race and ethnicity relations, wellbeing
  - Political science: government performance, government policy effectiveness, election, protests and social movements
  - Economics: measuring economic growth with big data
  - History: historical development of an idea
  - Psychology: measuring personality with big data
  - Communication and information science: content and spread of fake news/hate speeches on social media

Literature Review

- Short presentation: briefly discuss the idea and get feedback from instructor/classmates
- Written Report: each literature review report should contain 6 - 8 pages, 12 points, double space.
- Use this to prepare your final paper

Final Paper/Project

- By default, you need to write a research final paper
  - The intended audience for research final paper are other researchers
- Alternatively
  - You can build a website that has nice visualization of some social science concepts/measurements/dataset
  - or some real software that make people easier to use big data / machine learning
  - Project should attract layman
  - If you want to go this route, discuss with instructors early

Final Paper/Project

- If you choose to write an research paper:
  - Presentation (20 minutes): follow a standard presentation style for academic talks.
  - Final paper/project: 15 - 20 pages, 12 points, double space, including Tables, Figures and References.
- If you choose to do a project:
  - Presentation (20 minutes). Show case your project in front of the class.
  - Technical report: a short write up on short background/dataset/methods; 10 pages, 12 points, double space;

Roadmap

- We will discuss 10 characteristics of big data, following
- Chapter 2, Matthew Salganik, *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2019
- `https://www.bitbybitbook.com/en/1st-ed/observing-behavior/data/`
- After our discussions, you can critically evaluate pros and cons of big data

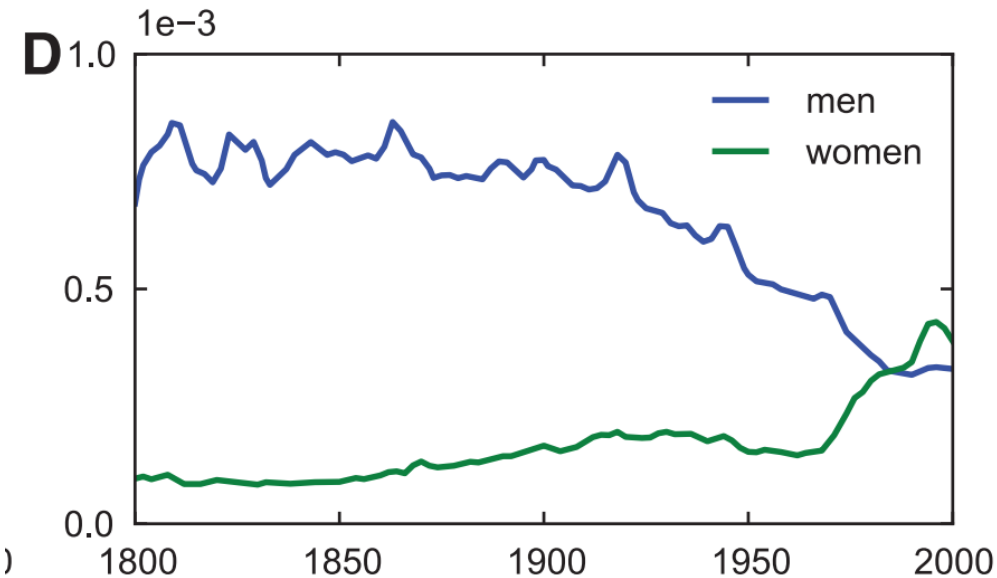
Characteristics 1: Big

- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden, *Quantitative Analysis of Culture Using Millions of Digitized Books*, Science **331** (2011), no. 6014, 176–182
- They turned Google Books into word counts and released the data

[Our] corpus contains over 500 billion words, in English (361 billion), French (45 billion), Spanish (45 billion), German (37 billion), Chinese (13 billion), Russian (35 billion), and Hebrew (2 billion). The oldest works were published in the 1500s. The early decades are represented by only a few books per year, comprising several hundred thousand words. By 1800, the corpus grows to 98 million words per year; by 1900, 1.8 billion; and by 2000, 11 billion. The corpus cannot be read by a human. If you tried to read only English-language entries from the year 2000 alone, at the reasonable pace of 200 words/min, without interruptions for food or sleep, it would take 80 years. The sequence of letters is 1000 times longer than the human genome: If you wrote it out in a straight line, it would reach to the Moon and back 10 times over.

Characteristics 1: Big

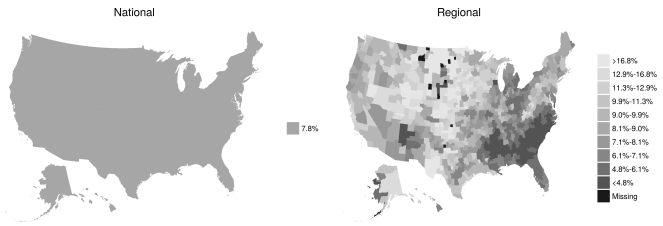
- Explore their project here: `https://books.google.com/ngrams`
- E.g., "In the battle of the sexes, the women are gaining ground on the men"



- [In class discussion]: do we really need this many data to draw the conclusion that women's right are rising? Can't we use smaller data to reach the same conclusion?

Characteristics 1: Big

- Big data are good at showing heterogeneity, which cannot be obtained by smaller data
- Chetty et al. 2014: estimates of a childs chances of reaching the top 20% of income distribution given parents in the bottom 20% .
- "The regional-level estimates, which show heterogeneity, naturally lead to interesting and important questions that do not arise from a single national-level estimate. These regional-level estimates were made possible in part because the researchers were using a large big data source: the tax records of 40 million people.



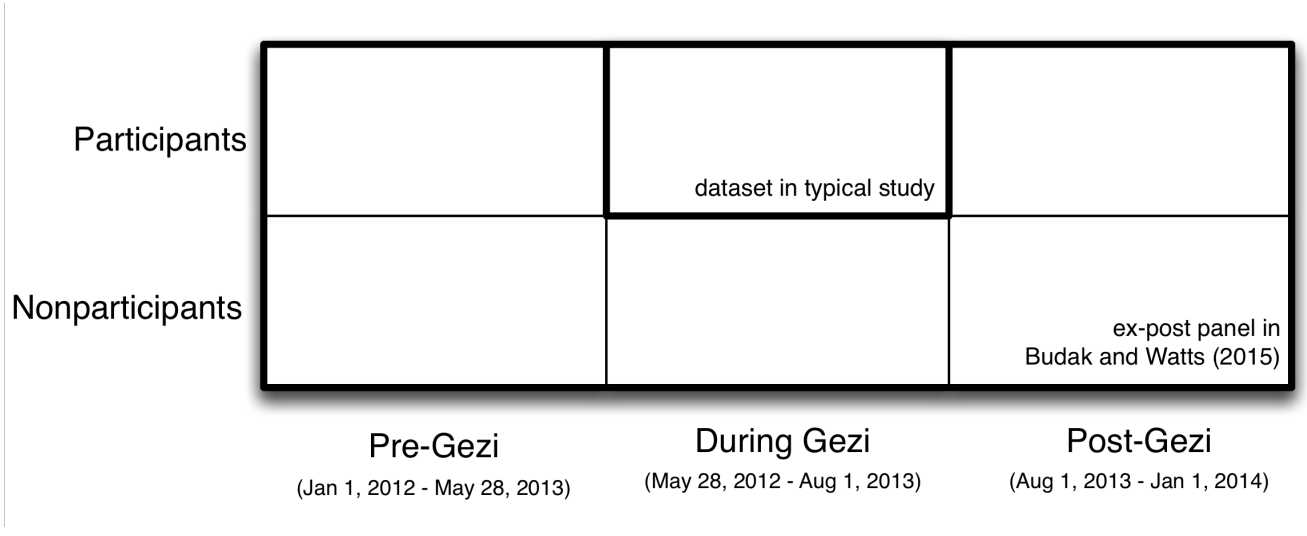
- https://www.nytimes.com/2015/05/04/upshot/

Characteristics 2: Always-on

- Traditional data survey: once a year, or on demand
- Big data: always-on measure
- Ceren Budak and Duncan Watts, *Dissecting the Spirit of Gezi: Influence vs. Selection in the Occupy Gezi Movement*, Sociological Science **2** (2015), 370–397
- What kinds of people were more likely to participate in the Gezi protests in 2013?
- Whether participation changed attitudes of participants and nonparticipants differently?
- Hard with survey data:
  - You cannot predict when a protest occur, and thus cannot get pre-protest information
    - If you ask people after protests, they may change their answer/memory based on the outcome of protests
  - It's also not easy to get samples of non-participants: selection bias

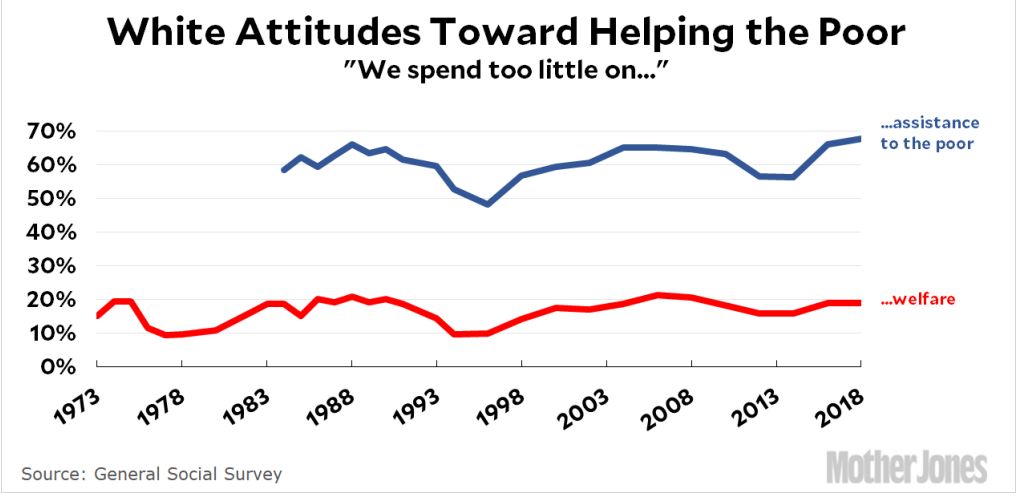
Characteristics 2: Always-On

- Using geolocated posts on Twitter



Characteristics 3: Non-Reactive

- Big data are mostly obtained unobtrusively;
  - People are generally not aware that their data are being captured
- Survey and lab experiments obtain data obtrusively, and results often depend on how you ask



Characteristics 3: Non-Reactive

- [In class activity]: is non-reactiveness always good?
- What people put on social media may be just showing off, not their daily lives
- Sometimes it is quicker to ask, especially for questions that are less likely to vary depending on how you ask

Characteristics 4: Incomplete

- "No matter how big your big data, it probably doesn't have the information you want"
- This is a property of re-purposing the data;
  - for survey and lab experiments, you can in principle ask what you want
- Three types of data are especially likely to be missing
  - demographic information
    - E.g., Google N-grams has gigantic dataset, but does not directly has author's biography
  - behavior on other platforms
  - data to operationalize theoretical concepts
    - people who are more intelligent earn more money
    - how do you measure intelligence with big data? Not easy

Characteristics 4: Incomplete

- But incompletness may also occur for traditional survey data.
  - A national representative survey does not
- [in class activity] E.g., a classical argument in social networks: the more centered you are in social network, the more wealthy you are
  - How do we test this argument with survey data?
  - How can we test this argument with big data?
  - What will be the best data source you can think of?

Characteristics 5: Inaccessible

- Many useful data are not directly available to researchers; they are stored in government and company servers
- Reason 1: commercial/government secrets
- Reason 2: terms-of-service agreements
- Reason 3: releasing data sometimes lead to privacy concerns

Characteristics 6: Nonrepresentative

- Many big data sources are not representative samples from some well-defined population
- [in class activity] So does it mean big data are not useful? When nonrepresentative data are useful?

Characteristics 7: Drifting

- Digital world changes so quick so that it's still too early to use big data to study long-term trends
- Population drift (change in who is using them)
- Behavioral drift (change in how people are using them)
- System drift (change in the system itself).

Characteristics 8: Algorithm Confounded

- Again, gov/companies control the data generating process
- E.g., to what extent your friends of friends are more likely to be your friends?
  - It's called *triad closure* in social network analysis
- With social networks such as Facebook, it's possible to empirical measure this quantity precisely
- Until Facebook began to recommend friends to users
- Now, are what we observe because of Facebook's recommendation or innate tendency for friends of friends to become friends?

Characteristics 9: Dirty

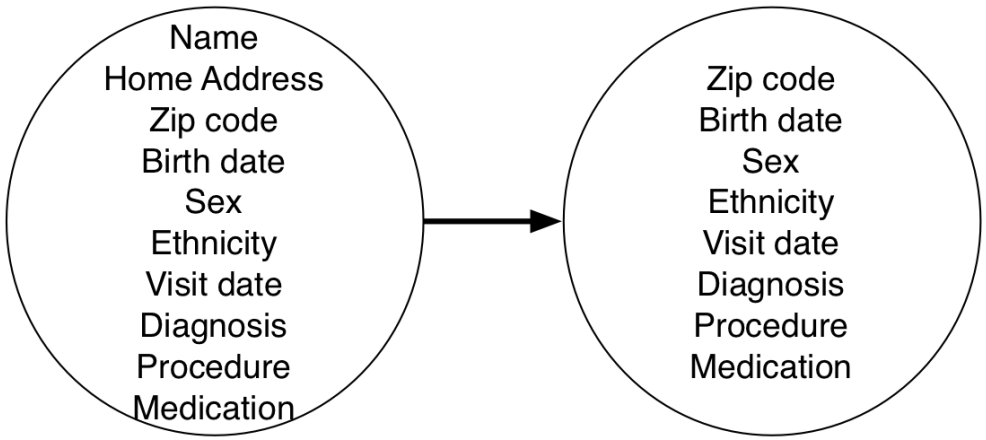
- "Big data sources can be loaded with junk and spam"
- Example: Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds, *Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution*, PLOS ONE **10** (2015), no. 10, e0137041
- Problem 1: OCR error
- The count of F-word between 1800 to 2000
- Are people suddenly become more polite after 1800?
- No! it's because *s* in old books are often written as a long s that looks like f before and around 1800s; so Google Books treat suck as the f-word in 1800.

Characteristics 10: Sensitive

- Some of the information that companies and governments have is sensitive.
- Even if we tried to anonymize data
- This lead to potential ethical questions

Characteristics 10: Sensitive

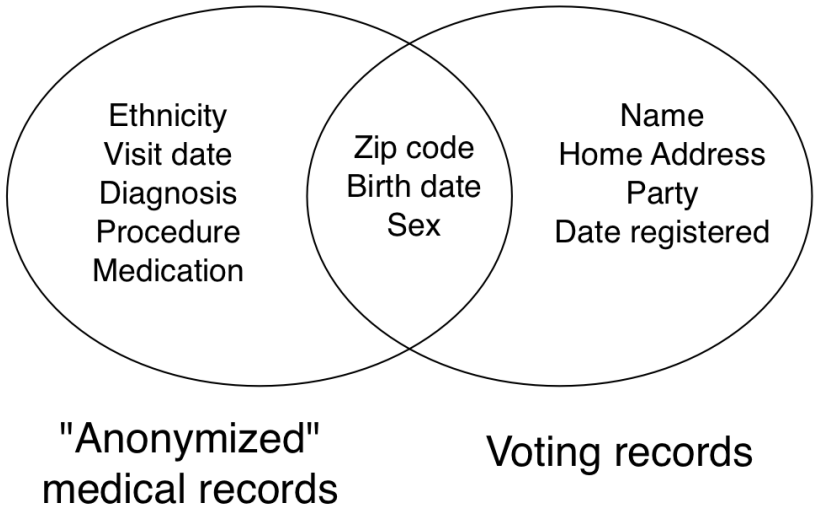
- Example 1: how anonymization fails
- Group Insurance Commission (GIC) was a government agency responsible for purchasing health insurance for all state employees in Massachusetts.
- GIC released some health information to spur research, and anonymized the part they thought were sensitive



"Anonymization"

Characteristics 10: Sensitive

- But Latanya Sweeney (now a Professor at Harvard) found that she could merge GIC data with public voter registration record



- And in this way, she was able to find a unique match: then governor of Massachusetts.
- Sometimes, even good intention and best effort to anonymize can lead to potential harm
- Things can only be worse if no effort has been made to protect privacy

Summary

- We have summarized 10 characteristics of big data
- You should be able to "describe the opportunities and challenges" brought by big data
- And when you evaluate future studies using big data
  - critically evaluate their strength and weakness
- I hope when you are writing your literature review and your final project, you can use some of these concepts to evaluate whether the datasets have these features/shortcomings

Task

- Imagine that I want to explore whether scientists are doing more research on "computational social science"
- I decide to use text data from arXiv, a place where scientists put their preprints before publication
  - certainly you can use other published article's electronic database, but most of which are not free so as a student you may choose the free one.

1. Use existing datasets

- Google
- Some collection of datasets:
  - Google Dataset search: `https://datasetsearch.research.google.com/`
  - Kaggle Dataset: `https://www.kaggle.com/datasets?tags=14104-text+data`
  - UCI's machine learning repository: `https://archive.ics.uci.edu/ml/datasets.php`
  - US patents: `https://www.google.com/googlebooks/uspto-patents-assignments.html`
  - Wikipedia texts: `https://en.wikipedia.org/wiki/Wikipedia:Database_download`

2. API: Application Programming Interface

- Easy; websites will provide you instructions to obtain some of their data
- Typically you need to register, obtain some key/password, and
- Some APIs are free while others are not.

API example: arXiv

- I will google "arXiv + API + python"
  - because I use Python; you can use other languages you like
  - if you google "arXiv + API + R", you can find R packages for similar functions

3. Web crawling

- Write a script to automatically download webpages/app data
  - In principle, you can download whatever you can see as a human
  - Technically more challenging; you need to have knowledge of basic HTML language and more python knowledge on handling computer network
  - Much more slower
  - Only use this when there is no API / API does not satisfy your need

Web scraping example: arXiv

- Step 1: download the HTML file
- Step 2: clean the data
- This is a tedious process and needs careful attention
- You can use browser's inspection tools
- From here, there is no single method that can help you; you need to use your own intelligence for each website

Some cautions

- static vs dynamic websites
  - arXiv is a static website which means that our returned HTML page already have all information you need
  - Most social media are dynamic website: URL is the same, content is different (depending on who logged in)
    - This creates another layer of complexity
- Desktop sites vs. mobile app

Caution: speed and fair use

- Most websites have a "robots.txt" that specify what you can crawl and the speed you can visit them
- e.g., `https://arxiv.org/robots.txt`
- Respect the website's requirement; otherwise your IP may be banned from access
- Be aware of legal risk; using API for commercial purposes can be risky