# SOSC 4300/5500: Overview

Han Zhang

Sep 8, 2020

# Outline

Computational Social Science

Logistics

Git and GitHub

Big Data: Opportunities and Challenges for Social Scientists

# 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

## 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.
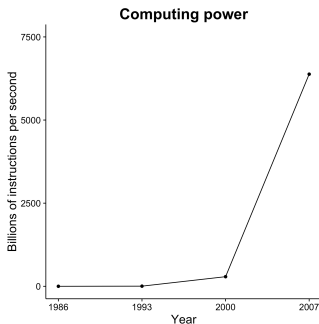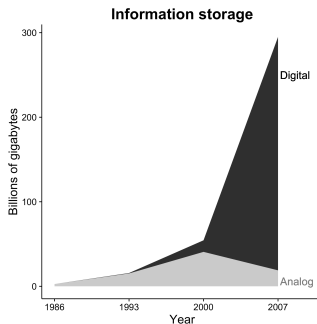
# Welcome to the digital age

- 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 <span style="color:red">regression</span> 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 <span style="color:red">machine learning</span>
  - 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)
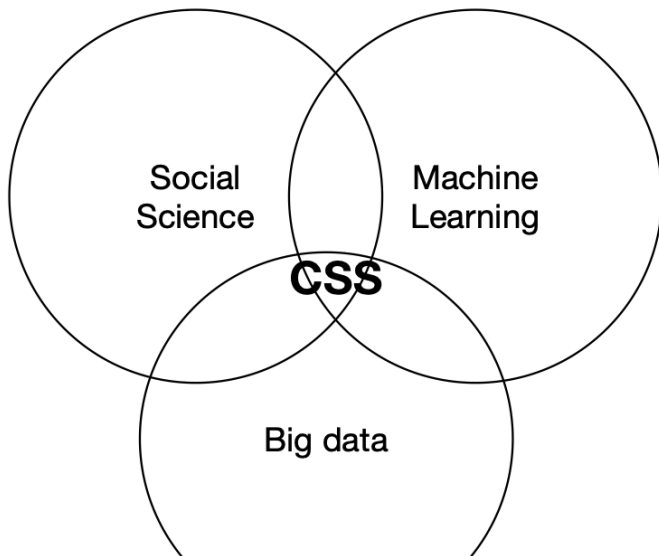
# Study Goals

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 (covered in Tutorials)
   - Getting data
   - Managing data
   - Analyzing data with appropriate methods

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: Tue 2-3PM (or email me to find a time)
    - In office (if you are on campus)
    - Zoom: `https://hkust.zoom.us/j/6522716568`

Teaching Assistant: PENG, Wenwei

- Office: 3001
- Email: wpengad@connect.ust.hk
- Office Hour: TBD
    - Zoom: `https://hkust.zoom.us/j/8653678438`

## Online teaching

- Some rules:
    - Keep the video on (this will be counted as class participation)
    - Mute yourself while not speaking
    - Feel free to stop me at any time if you have any question
    - Recordings will be available after class

## Course material

- We will use GitHub for teaching
- 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 | Topic |
|------|-------|
| 1 (Sep 8) | Introduction; big data |
| 2 (Sep 15) | Machine learning; prediction vs explanation |
| 3 (Sep 22) | Survey |
| 3 (Sep 22) | Text (I) |
| 5 (Oct 6) | Text (II) |
| 6 (Oct 13) | Text (III) |
| 7 (Oct 20) | Images, Videos and Audio |
| 8 (Oct 27) | Network (I); |
| 9 (Nov 3) | Network (II); |
| 10 (Nov 10) | Network (III): |
| 11 (Nov 17) | Causal inference: online field experiment |
| 12 (Nov 24) | Causal inference: big data |
| 13 (Dec 1) | Presentation of final paper and projects; wrap up |

## Grading Components

|                              | UG students        | PG students        |
|------------------------------|--------------------|--------------------|
| Attendence and participation | 10%                | 10%                |
| Homework assignments         | 20% (independent)  | 20% (independent)  |
| Literature review            |                    |                    |
| Report                       | 15% (3-4 people)   | 15% (1-2 people)   |
| Presentation                 | 10% (3-4 people)   | 10% (1-2 people)   |
| Final Paper/Project          |                    |                    |
| Presentation                 | 15% (3-4 people)   | 15% (1-2 people)   |
| Write-up                     | 30% (3-4 people)   | 30% (1-2 people)   |

## Class participation

- 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.

# Literature Review

- Select a research topic and summarize how researchers have used computational methods and/or big data to study this particular research area.
- 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

- Your performances will be accessed in two ways
    - Written Report: each literature review report should contain at least 10 pages, 12 points, double space. Spell out clearly contributions of each group member in the first page of your report.
    - Presentation of literature review (15 minutes): each student/group needs to present their literature reviews in class.

# Final Paper/Project

- You can choose to write a research final paper
  - The intended audience for research final paper are other researchers
- Or a project that analyze a "real-world" case.
  - Project should attract layman

# 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: 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;

## We will use GitHub for teaching

- We will use GitHub for teaching
- GitHub is commonly used among researchers and companies
- It allows us to manage individual research workflow more easily
- And allows team work easily (especially if you have lots of codes/data)
- Last, GitHub simplifies sharing data and codes, making research more transparent and useful for the community

## So what is it?

Linus Torvalds: creator of Linux

## When Linux grows

- Over 30,000 files; 15 Million lines of codes in 2012
- Version control
  - We will update a file, but want to keep some old version if we need to roll back
  - Renaming files to file.r, file_ver1.r, file_ver2.r $\cdots$: an ugly solution
- Project management:
  - Multiple people working on the same file
  - Contributors send Linus code chunks, and Linus put these code chunks to its appropriate place manually
  - Error prone: over 3,500 lines of codes added per day
- So Linus intended Git

## Git for version control

- Version control (If you are working individually):
    - add file.r=: start to track changes of file.r
    - commit file.r= : record changes locally
        - First commit records the oldest version of a file
        - Newest commit records the most recent version of a file
        - Only changes in a file will be saved; saving disk spaces
    - push file.r= : send committed changes to GitHub so that others can view it

## Git for version control: example

- Say you are working on assignment 1, with a R script
  homework1.r
- Step 1: add homework1.r to tell git to pay attention to
  changes in this file
- Whenever you want to save a version for backup, commit
  homework1.r
- When you feel you are ready to submit, push to GitHub
  - Of course you can push multiple times, if deadline is not
    passed.

## Git for project management

- Project management for groups: (called repository or repo)
  - Owner controls master branch; this is the central and authoritative version
  - Every member works with his own local version (called local branch)
  - Only owner can directly push to the central version
  - Members can create pull request, which allow the owner to *pull* changes from local branch to master branch

## GitHub

- Git is free;
- GitHub is an online service that stores Git repository;
    - Student can get a free version, allowing you to obtain private repo.
- GitHub also allows us to share codes and data easily
- https://github.com/HKUST-SOSC4300-5500/ Tutorial-Material/blob/master/week1/ 1-Hello-World.ipynb

# Three ways to use Git

- Command line: for advanced and interested users
- Recommended: Through GitHub's official app. MacOS and Windows version can be downloaded at `https://docs.github.com/en/desktop/` `installing-and-configuring-github-desktop/` `installing-github-desktop`
- Directly upload files on GitHub website
  - Note: if you directly upload a file on GitHub in your own repo, you are performing `add`, `commit` and `push` operation simultaneously

# How we use GitHub for this class

- Lecture and tutorial material will be available at
  `https://github.com/HKUST-SOSC4300-5500`
  - You can view and download the files
- Assignment will be submitted on GitHub:
  `https://classroom.github.com/classrooms/`
  `70853257-hkust-sosc4300-5500-classroom-1`
  - Homework will be <span style="color:red">private</span>: only you and instructors can see your commits
  - Literature review and final paper/project will be <span style="color:red">public</span>: since these knowledge will be beneficial to others
  - We will set deadlines; your last `push` before the deadline will be automatically treated as your final submitted version

## How we use GitHub for this class

- You can ask question about lecture or tutorials under "Issues" tab
- TA and I will answer these questions

# Computational Social Science (CSS)
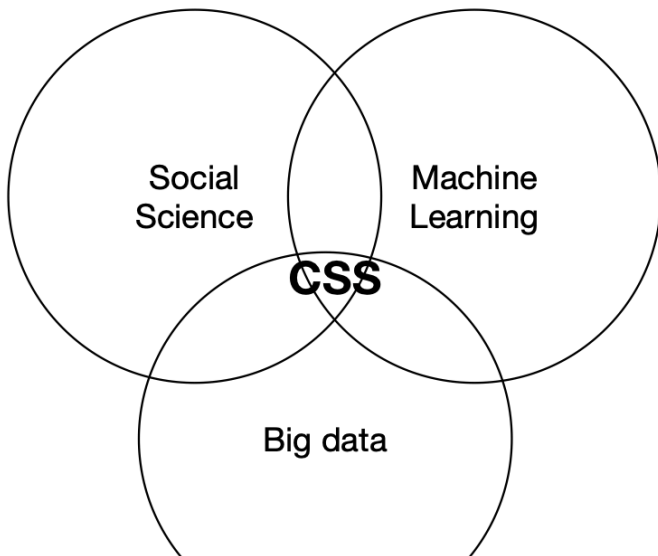
# Roadmap

- 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
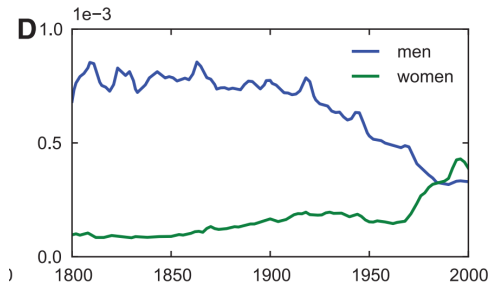
# Characteristics 1: Big

- 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.
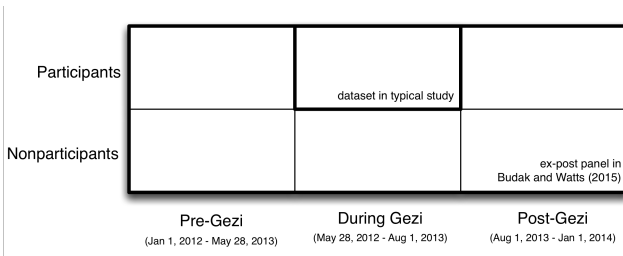
# Characteristics 2: Always-on

- 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
  - It's also not easy to get samples of non-participants

# Characteristics 2: Always-On

- Using geolocated posts on Twitter



|  | Pre-Gezi<br>(Jan 1, 2012 - May 28, 2013) | During Gezi<br>(May 28, 2012 - Aug 1, 2013) | Post-Gezi<br>(Aug 1, 2013 - Jan 1, 2014) |
|---|---|---|---|
| Participants |  | dataset in typical study |  |
| Nonparticipants |  |  | ex-post panel in<br>Budak and Watts (2015) |

# 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



**White Attitudes Toward Helping the Poor**
"We spend too little on..."

Source: General Social Survey
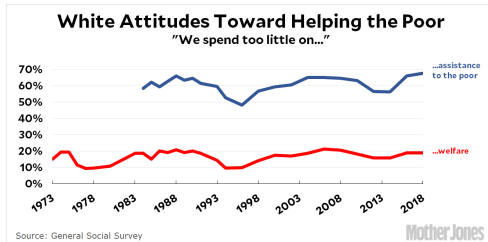
# Characteristics 3: Non-Reactive

- [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 extend 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.
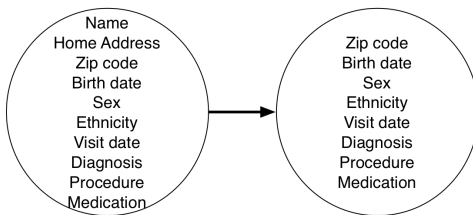
# Characteristics 9: Dirty

- Example 2: figure vs Figure
- Why Figure is significantly used more than figure?
- Oversampling of scientific literature

# 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
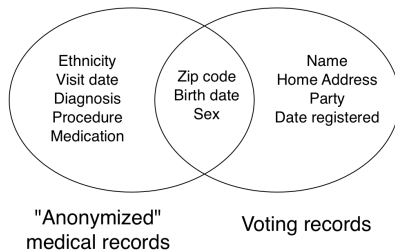
# Characteristics 10: Sensitive

- 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



Name
Home Address
Zip code
Birth date
Sex
Ethnicity
Visit date
Diagnosis
Procedure
Medication

Zip code
Birth date
Sex
Ethnicity
Visit date
Diagnosis
Procedure
Medication

"Anonymization"

## Characteristics 10: Sensitive

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



"Anonymized" medical records — Ethnicity, Visit date, Diagnosis, Procedure, Medication | Zip code, Birth date, Sex | Voting records — Name, Home Address, Party, Date registered

- 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

# Next week's plan

- We will cover basic ideas of machine learning
- It helps you to "describe the opportunities and challenges" brought by machine learning
- And critically evaluate strength and weakness of using machine learning to study social phenomena