Slide 1:

Computational social science is just a hot term out there. It is exciting due a lot of opportunities but also a lot of challenges. Today I am just giving a brief overview about what is big data in general, and what we can do with big data. Each topic could be a class for a whole semester. I am not expecting you nail all the terms, all the methods that I mention later. But I do want you to be more familiar with some of the terms.

Slide 5:

The manifesto of Computational Social Science. If there is one book I would suggest everyone in social science to read about big data or computational social science, I will recommend this one.

Slide 6:

Bigness: after the entering the digital era, the storage of data, and computing power has increased exponentially. How big would you consider big? The term itself is quite fuzzy to some extent since everyone is using it. Big Data, it is commonly stated, far more than an average desktop computer can provide, with the data often stored in the cloud across several servers and locations. However, when we are using the term, it was not quite big. Be cautious when you are using the term. Sometime because it means it requires more computing power than usual.

Always-on: Always-on big data enables that the study of unexpected events and real-time measurement (social movement; social media usage during public crisis; traffic mobility change based on your GPS in your phone, it is good measure of lockdown during COVID-19)

Nonreactive: reactivity: one challenge of social research is that people can change their behavior when they know that they are being observed by researchers in a survey or experiment. The realness, or then unobtrusiveness measurement in big data sources is much less likely to change behavior

Slide 12:

Classification (DV is binary variable, text – spam or not spam; image – dog or cat) and regression (DV is a continuous variable, stock market price).

Unsupervised: let the data speak itself: Topic modeling is a way to find hidden patterns in a large collection of documents or texts.

Supervised machine learning need training data to learn the pattens. That is why we need annotation or labelled data. You trained it, you tested it and you applied it to the rest. The large language model, or imagenet, I believe we all have been the annotators – when you guys log in some website, the website tests you whether you are a robot or not.

Slide 13: Capcha test

Slide 15:

* Content analysis: codebook – good training data
* A zero-shot model is a machine learning model that can recognize objects or categories it has never seen before, without any additional training or examples. It is like a person who can recognize a new type of animal they have never seen before, just by using their knowledge of other animals and their characteristics.
* Fine-tuning. A one-shot model is a machine learning model that can recognize objects or categories with only one example or training example. It is like a person who can recognize a specific breed of dog after seeing it only once, by using their knowledge of other dog breeds and their characteristics.
* GPT
* We pay less attention into developing the algorithm but some scholars also start to get involved.
* Explanatory modeling (tree-based modeling, extract the most important feature:
  + Statistical model is careful operationalization of theory
  + Parameter estimates are the objects of interest
  + Statistical model itself is the object of interest.
  + Variance less important than bias
* Predictive modeling – forecasting (CDC: COVID-19)
  + Objects of interest are the variable values.
  + Statistical model chosen to produce best predictions.
  + Model does not necessarily correspond to any theory
  + Prediction is typically prospective
  + Variance possibly more important than bias

Slide 25:

Bidirectional Encoder Representations from Transformers is a transformer-based machine learning technique

Severe toxicity: This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.

Slide 25:

Each score was ranked between 0 and 1, we have computed scores for each single tweet and we aggregate on the hour. As we can clearly see that after Trump’s tweet, there are way more tweets from the very first time-series plot. So we also average down the toxicity based on the number of tweets in the hour as well. We can see across the six types of toxicity, there are more fluctuation before Trump’s tweet, and the toxicity volume remain comparatively consistent after the treatment.

Slide 26:

Interrupted time- series analysis is a powerful quasi-experimental design for assessing the longitudinal impact of an event or intervention. Beside the treatment effect, we are also interested in the so-called pre-treatment trend, as the the change in the volume of toxicity in the dataset per unit time increase. And also the so-called slope change.

Slide 27:

As a quick sum-up, the Trump effect only existed in three types of toxicity:

Severe toxicity: A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. Normally, when user use ’Chinesvirus’ it tends to be toxic. That is why not many differences before and after Trump’s tweet. However, ‘Trump effect’ still exist for the severe toxicity.

Threat: Describes an intention to inflict pain, injury, or violence against an individual or group.

Identity attack: Negative or hateful comments targeting someone because of their identity.

Only threat detects a lasting rise caused by Trump effect, as indicated by the positive shift instantly following Trump’s tweet, and the positive post-treatment slope change

The joint effect of the quadratics terms only worked for severe toxicity, which is slowly progressively increase over time

But the pre-treatment trend and the slope change might vary.

Slide 29: JOA, IJOA, PRR

IBM’s personality insight engine

Slide 32: Image-as-data or video-as-data: nonverbal communication. Attention in german parliement

Audio-as-data: political debate: passionate speech, tedious monologue.

Network/Graph-as-data: Node classification; Link prediction

Slide 33: Personality traits and risk perception and opinion expression

Slide 36: Misinformation

Slide 38:

Incomplete: No matter how big you data is, it probably doesn’t have the information you have (ban by twitter)

Inaccessible: Data held by companies and governments are difficult for researchers to access. Twitter in general more friendly to researcher. Facebook is hard to work with – their own portal and VPN, and they don’t really to share data. Not to mentioned TikTok (Bytedance, the parent company) and now Twitter

Nonrepresentative: Nonrepresentative data are bad for out-of-sample generalizations, but can be quite useful for within-sample comparisons. Can Twitter represent America? Not at all! Only the 22% of American adults are using Twitter. Twitter users are younger, more likely to identify as Democrats, more highly educated and have higher incomes than U.S.

Drifting makes it hard to use big data sources to study long-term trends. Population drift (change in who is using system, for instance, scholars found during the 2012 US presidential election, the proportion of tweets about politics that were written by women fluctuated from day by day), usage drift or behavior drift (for instance, the hashtags might change when protest evolves during a social movement), and system drift – the system itself, engineering speaking has changed. For instance, the word length you can tweet, extending from 140 to 280 characters. It is tied with the next characteristics.

Algorithmically confounded: Behavior in big data systems is not natural; it is driven by the engineering goals of the systems. Facebook has changed their algorithm about what you can see on your own feed, and weight more on specific emotions (anger) and not based on timeline.

Slide 39:

Dirty and noisy: a lot of junk and spam. The comments you guys collected from social media, a lot of acronyms, abbreviations, slang, emojis, incorrect grammar. Data scientists are spending a lot of times in data cleaning.

Sensitive: some of the data is pretty sensitive and might cause potential harm. Netflix has a prize to call for improvement of their recommendation system. Although they anonymized the personal identification, researchers are still able to retrieve the information with external dataset (for instance, zip code, birth date, sex).

Slide 40:

Another case is the Parlor app – here is all the geo-locations that have uploaded video or image during January 6th Capital hill riots. Household address and attack.