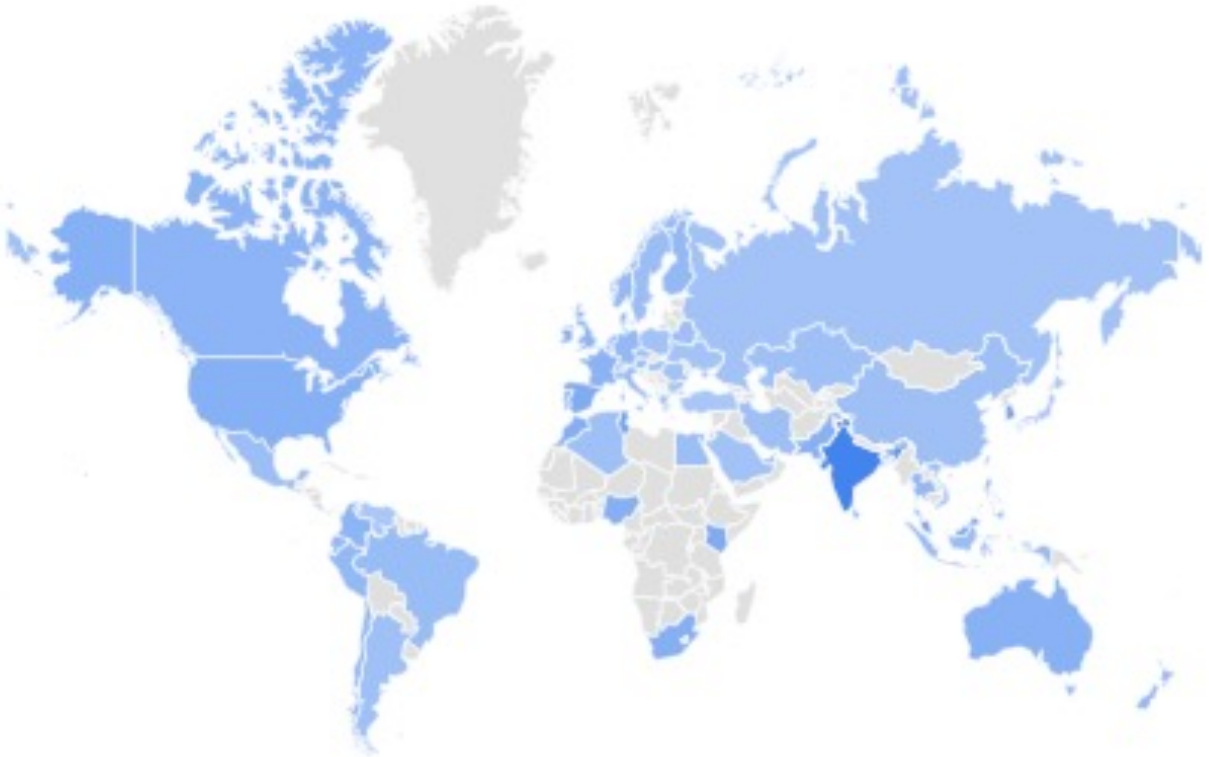
Fall 2022
PS0700 Research Method in Political Science

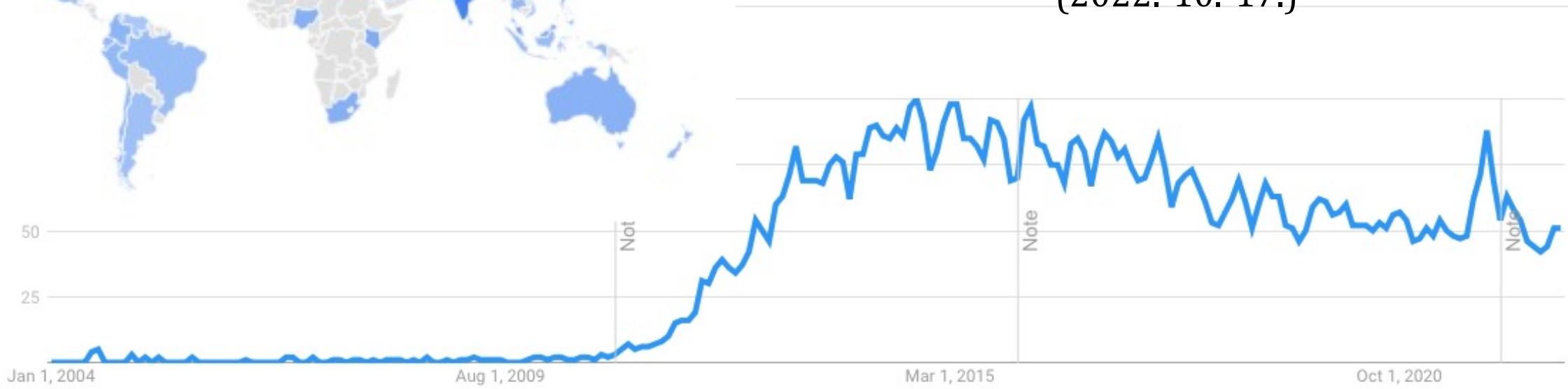
Aggregate Analysis, Content Analysis and "Big Data"

Jungmin Han

Prevalent usage of “Big Data”



Google Trend “Big Data”
(2022. 10. 17.)



The Outline of Today's Lecture

- What is "Big Data"?
- Opportunities of Big Data in Political Science
- Limitations of Big Data in Political Science

Part 1.

What is “Big Data”?

What is “Data” ?

- *Data is “Information, especially facts or numbers, collected to be examined and considered and used to help decision-making” Cambridge Dictionary*

- *Types of Data:*

1) *Purpose: Is data designed to fulfill a specific purpose?*
   - *“Designed/Made” vs. “Organic/Found”*

2) *Structure: Is the format of data organized and structured?*
   - *“Structured ” vs. “Unstructured”*

Data Type (1): *"Designed/Made" vs. "Organic/Found"*

- *"Designed/Made" D*ata
  - Data created for research purposes or designed to answer specific questions
  - More directly relevant to research questions or concepts
  - Measurement validity is usually higher
  - e.g. Survey Data; Interviews; Experiment Data

- *"Organic/Found" D*ata
  - Data created a by-product of another process of activity
  - Less relevant originally to research questions or concepts
  - Measurement validity needs to be assessed
  - e.g. Friends lists on social media; Online browsing and purchasing history

Data Type (1): *Example of "Designed/Made" Data*

WVS 2017-2021: WAVE 7

2017 -2021 WORLD VALUES SURVEY WAVE 7

MASTER SURVEY QUESTIONNAIRE

ANES 2020 Time Series Study

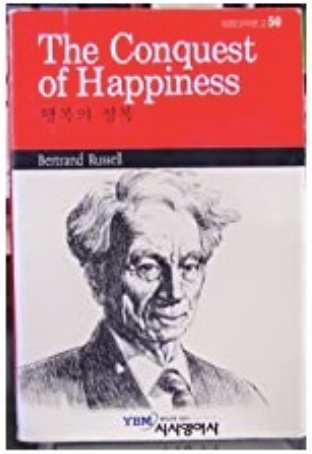
Pre-Election and Post-Election Survey Questionnaires

July 19, 2021

Data Type (1): *Example of "Organic/Found" Data*

Your Items

Saved for later (41 items) Buy it again



The Conquest of Happiness (In English and Korean)

Currently unavailable.

Delete

Add to list

DreamWorks Girls' Spirit Riding Free Top & Leggings...

$16.05

Only 5 left in stock - order soon.

prime & FREE Returns

Size: 5

Move to cart

Delete

Horse Girl Temporary Tattoos Birthday Party Supplies Dec...

$8.99

In Stock

prime & FREE Returns

Move to cart

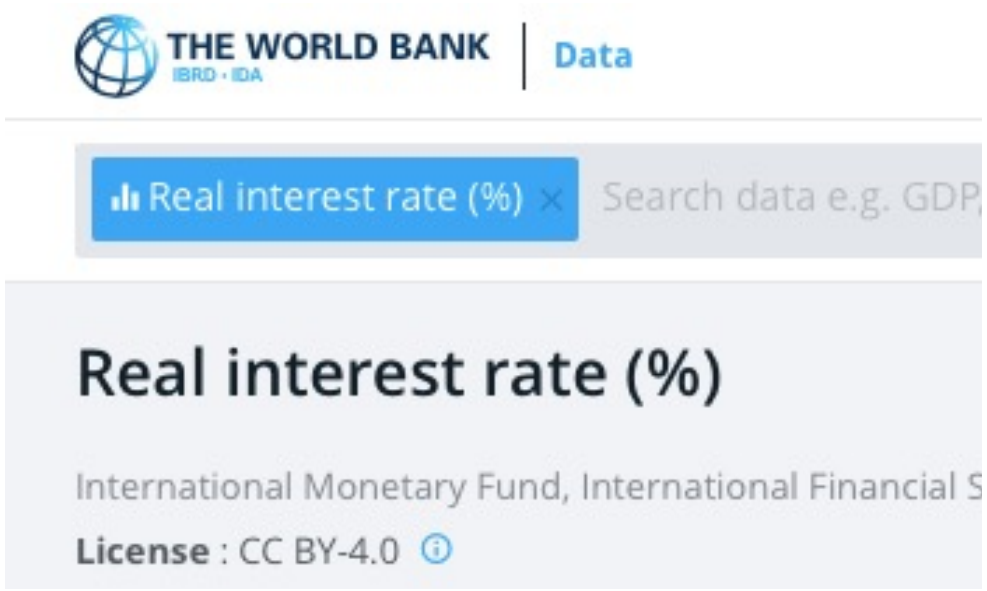
Delete

Add to list

Data Type (2): "*Structured" vs. "Unstructured"*

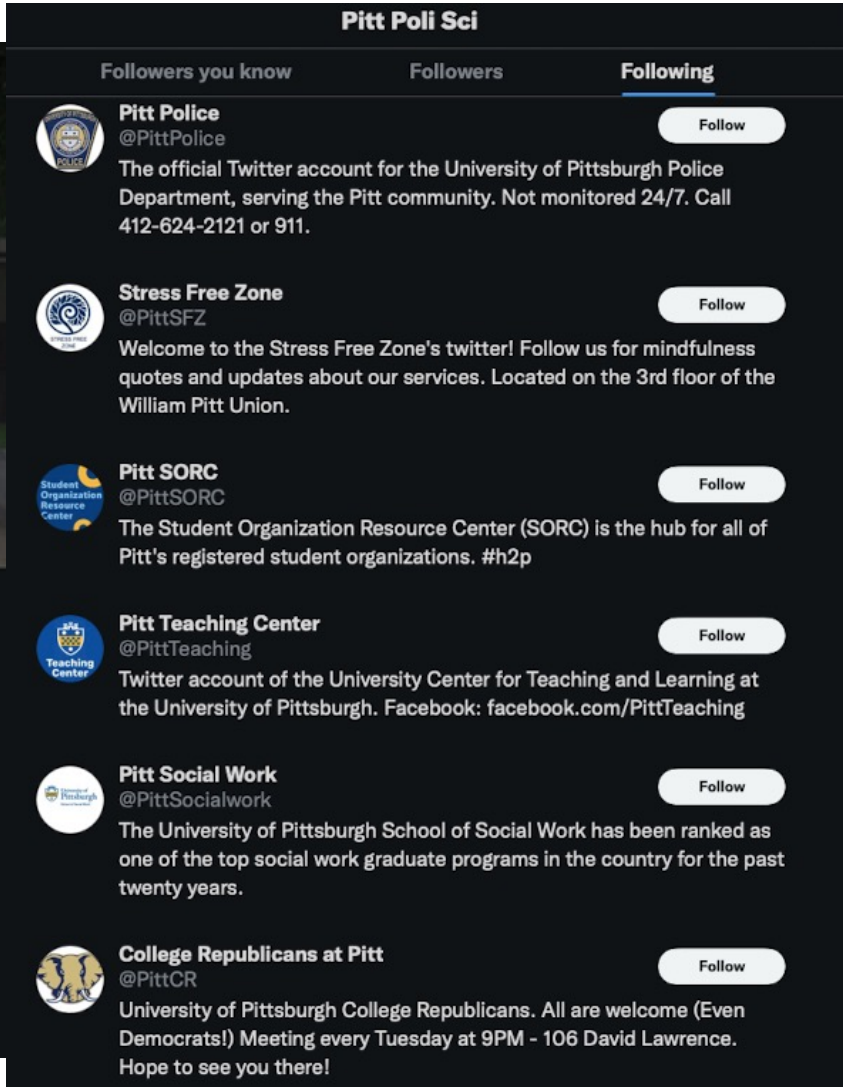
- *"Structured" Data*
  - Information is stored neatly in rows (observations) and columns (variables)
- *"Unstructured" Data*
  - Information is stored "free-form" in texts, images, or audio-visual format etc.
  - Rich in information, BUT needs lots of processing to extract the information out before usable for analysis
  - In social and political science, unstructured textual data is commonly used
    - E.g. Different linguistic styles between political leaders (Jordan et al., 2019; Schoonvelde et al., 2019); Predict violence onset from newspaper texts (Mueller and Rauh, 2017)

Data Type (2): Example of "*Structured* " *Data*



Data Source	World Development Indicators						
Last Updated Date	2022-09-16						
Country Name	Country Code	Indicator Name	2010	2011	2012	2013	2014
Aruba	ABW	Real interest rate (%)	11.658484092447	4.79578773803402	8.21727796856164	10.6993243243243	4.85160128102481
Africa Eastern and Southern	AFE	Real interest rate (%)					
Afghanistan	AFG	Real interest rate (%)	11.3640937400601	-1.24150600219686	7.17438724378044	9.78449593802676	14.3516893399556
Africa Western and Central	AFW	Real interest rate (%)					
Angola	AGO	Real interest rate (%)	-6.94413460977478	-9.87814512128374	8.76259485211296	12.6058526338601	12.3777859689274
Albania	ALB	Real interest rate (%)	7.97103659621924	9.88846223555399	9.73611207346728	9.51039749125922	6.32455753333484
Andorra	AND	Real interest rate (%)					
Arab World	ARB	Real interest rate (%)					
United Arab Emirates	ARE	Real interest rate (%)					
Argentina	ARG	Real interest rate (%)	-8.56602872006595	-7.77340523607891	-6.74819555440488	-5.48583918237899	-11.6006987702161
Armenia	ARM	Real interest rate (%)	10.6076357898925	12.9197578932998	11.2811932906832	12.2168430436348	13.7812486601176
American Samoa	ASM	Real interest rate (%)					
Antigua and Barbuda	ATG	Real interest rate (%)	9.3876761408781	9.79953139854016	7.95945311994109	11.0030464701126	8.0072745506121
Australia	AUS	Real interest rate (%)	6.04335497000121	1.44126323003352	5.09471889327147	6.34143469956045	4.44409935543742
Austria	AUT	Real interest rate (%)					
Azerbaijan	AZE	Real interest rate (%)	6.30070170030242	-2.88597938752249	15.0296068894244	17.6862204009871	19.3916777101463
Burundi	BDI	Real interest rate (%)	3.55415181609771	4.49248695285009	0.0226856189418669	6.6643715965697	9.84057048451342
Belgium	BEL	Real interest rate (%)					
Benin	BEN	Real interest rate (%)	4.19549019095698	1.46803081679031	-2.46960852040217	3.83659295370063	5.5999917955037
Burkina Faso	BFA	Real interest rate (%)	1.28514116887426	-1.3885289905265	-0.741639727368065	7.57802019961703	5.99591815146962
Bangladesh	BGD	Real interest rate (%)	4.7361235719975	5.06419768357502	5.34333265724941	5.98869390785248	6.88586627498441
Bulgaria	BGR	Real interest rate (%)	10.3014064810926	4.25629214173815	8.50745697999523	8.96499492994047	6.83032456491182
Bahrain	BHR	Real interest rate (%)	-0.180215875919306	-2.68453373074554	2.93184586033418	5.52961851204117	7.66732842513014
Bahamas, The	BHS	Real interest rate (%)	5.91550048070293	5.97260932490227	1.43580032378823	3.26981245370623	0.830597823616353

Data Type (2): Example of "*Unstructured*" *Data*



Data

that is TOO BIG to design or structure (?)

What is Big Data? (cont.)

- Francis Diebold (2000), "'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting"

  "*Recently, much good science, whether physical, biological, or social, has been forced to confront—and has often benefited from—the "Big Data" phenomenon.* ***Big Data refers to the explosion in the quantity (and sometimes, quality) of available and potentially relevant data****, largely the result of recent and unprecedented advancements in data recording and storage technology*."

- O'Reilly (2005), "What is Web 2.0?"

"Big Data refers to ***a large set of data*** *that is almost impossible to manage and process using traditional business intelligence tools."*

What is Big Data? (cont.)

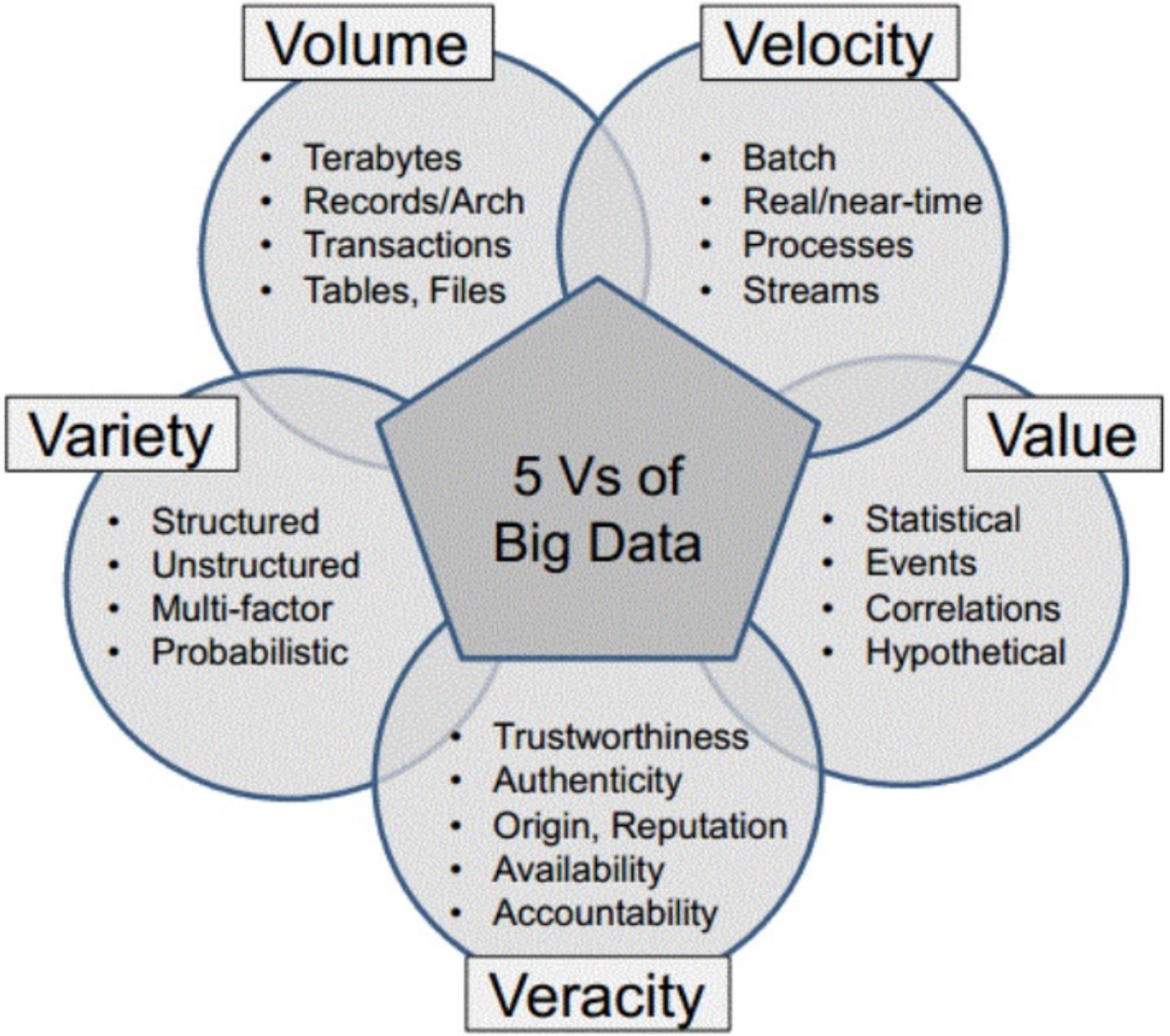
- Laney (2001), "3D Data Management: Controlling Data Volume, Velocity, and Variety."

  *"**High-volume, high-velocity and/or high-variety information assets** that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."*

- **Size, complexity, and technological challenges** provide one definition of big data (National Research Council 2013, Ward & Barker 2013), but they do not seem a sufficient basis for heralding a sea-change in our data environment.

- Beyond the sheer amount of data, the truly distinguishing features of the big-data revolution are the **new technologies for recording, connecting, networking, and creating information** (Brady, 2019).

"Five V" Characteristics of Big Data (*Hadi et al. 2015*)



- ***Volume***: *quantity of data collected*
- ***Velocity***: *speed of data collection*
- ***Variety***: *source, type, and context of data collected*
- ***Value***: *spectrum of modeling and prediction*
- ***Veracity***: *data quality and accuracy*

- *Variability, Volatility, and Visibility etc.*

Part 2.

Opportunities of “Big Data” in Political Science

Big Data in Political Science Research

- Promising area where big data can be used (examples)
  - Collective action and social movements
  - Political campaigns, mobilizations, and participations
  - Communication and information flow
  - State-citizen interactions

- ***How to use big data as a source of information to analyze political phenomenon?***
  (1) Individual behaviors and attitudes
  (2) Elite or state behaviors
  (3) Network Analysis

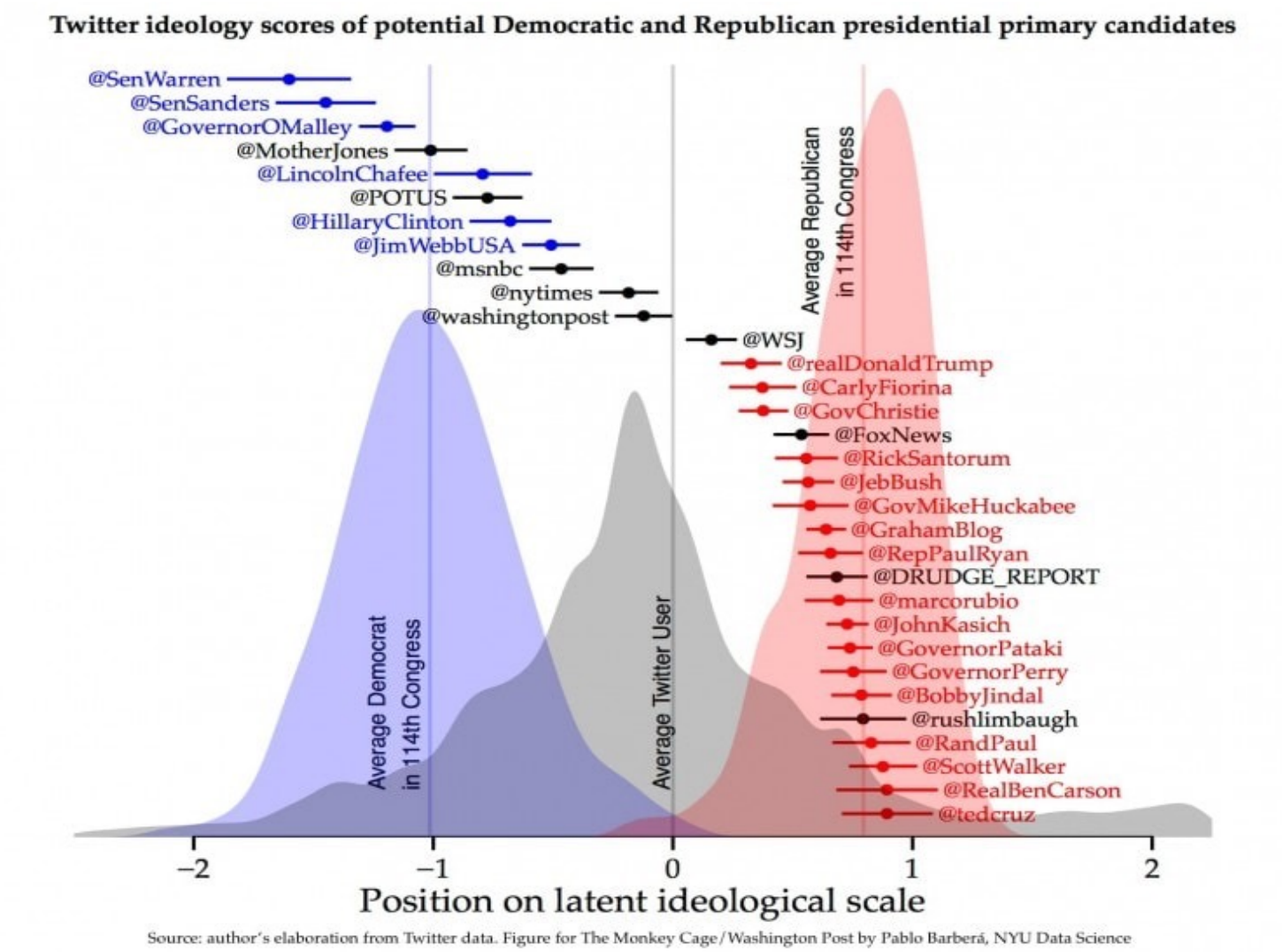
Big Data in PS 1: Infer Individual Attitudes and Behaviors

- Survey data is collected using direct questioning, and its data quality typically rely on:
  - Accurate introspection when asking about latent traits such as ideology and personal traits
  - Accurate reporting when asking about socially undesirable behaviors

- Big data provides *non-intrusive measurement* of behavior and public opinion (Nagler and Tucker, 2015)
  - Revealed preferences circumvents the problems associated with self-reporting

- Examples
  - Using Facebook likes to infer sexual orientation, personality etc. (Kosinski et. al 2013)
  - Using Twitter follow as networks to infer ideology (Barberá 2015)
  - Using Google searches to infer racial animosity (Stephens-Davidowitz 2014)

Big Data in PS 1: Infer Individual Attitudes and Behaviors



Barberá et al. (2015), Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?

Big Data in PS 2: Infer Elite or State Behaviors

- Elite or state behaviors are traditionally difficult to observe
  - Difficult to survey
  - Direct observation (e.g. interviews) are susceptible to misrepresentation
  - Many behaviors by nature are hidden (unobservable)

- Example:
  - Using social media posts to examine government censorship in China (King et al., 2013)
  - Observe how congressional members interact with constituents on Facebook (Pew Research Center, 2018)

Big Data in PS 2: Infer Elite or State Behaviors (Example)

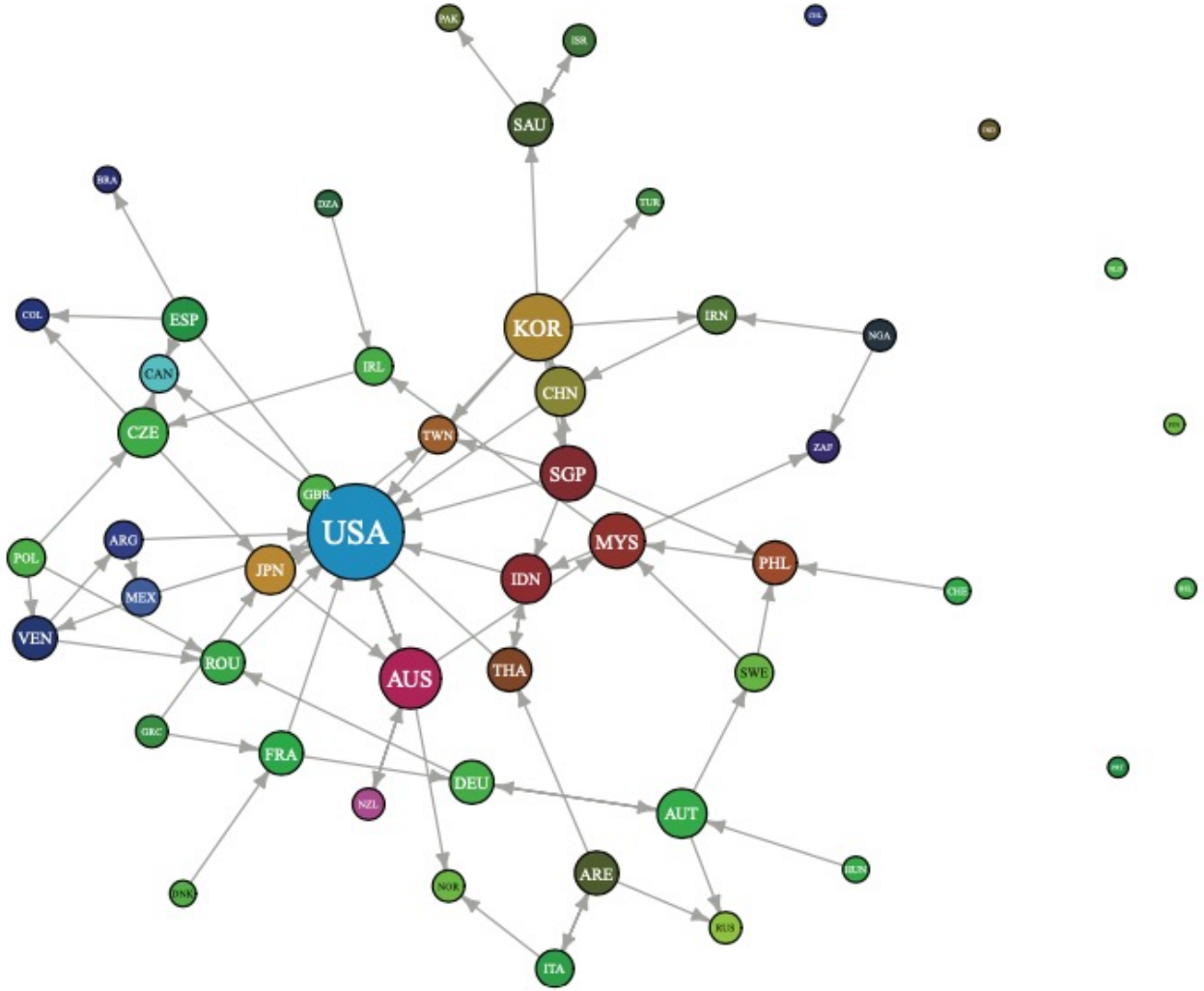
How Censorship in China Allows Government Criticism but Silences Collective Expression (King et al., 2013)

Big Data in PS 3: Network Analysis

- Network structures are difficult to measure by using survey data
- Social networks and online database provide an easy way to map network structure and density

- Examples
  - Using sharing and linking on Twitter to measure network pattern and influence of media outlets (Faris et al. 2017)
  - Analyzing international conflicts and cooperation by using a corpus of about 30 million media reports from about 275 local and global news sources (Minhas et al. 2016)

Big Data in PS 3: Network Analysis (Example)



Minhas et al. (2016), A new approach to analyzing coevolving longitudinal networks in international relations

Part 3.

Limitations of Big Data in Political Science

Main Limitations of Big Data in Political Science

(1) Non-representativeness

(2) Interpretation

(3) Measurement Validity

(4) Causal Inference

(5) Data quality

Limitation 1: Non-representativeness (1)

- Different big data sources have different user base
  - Different socio-demographic characteristics are correlated with usage and membership on different platforms and services
  - Non-random sample poses problems of generalizability (Nagler and Tucker 2015)

- Sample representativeness and its impact on external validity is always assessed in relevance to the target population we are trying to generalize to
  - Target population is not always “the population” (everyone on the Earth)
  - Sample from big data or social media is not always problematic
  - Depends on the research goal and question

Limitation 1: Non-representativeness (2)

- Platforms structure behavior patterns, not just recording them
  - Networks on Twitter is directional (one-sided follow)
  - Networks on Facebook is bidirectional (two-sided follow)

- Platforms encourage certain behaviors, potentially distorting the “true” pattern
  - Google auto-complete searches
  - Twitter suggests people for you to follow
  - YouTube recommends videos to watch

Limitation 1: Non-representativeness (3)

- Behaviors recorded in big data = behaviors in real life?
  - Some are records of offline behaviors or behaviors with offline analogous
    (e.g. voting records, purchase pattern, media exposure and consumption)
  - But, many are behaviors that only exists online, or specific to platforms, that do not have an offline analogous
    (e.g. retweets, likes)

- Are online behaviors representative and generalizable? (Jungherr & Yannis 2017)
  - Is a specific type of online behaviors a good indicator for the concept of interest?
    (e.g. Retweet = Political Participation ?)
  - Do online behaviors reflect offline behavior?
    (e.g. Civility in Online = Civility in Offline ?)

Limitation 2: Measurement Validity

- Is the indicator capturing the concept that we are interested in?

- Which indicator is a valid measure of a user's "influence" on Twitter?
  - Number of followers? Number of retweets? Centrality in network?

- Which indicator is a valid measure of "new consumption" on Facebook?
  - Headline on Timeline? Shared? Clicked?

Limitation 3: Interpretation

- What's in the text?
  - Sincere or sarcastic?
- What's in a retweet?
  - Endorsing? Mocking? Hate-retweeting?
- What's in a search query?
  - If we observe increase in Google search trend for "impeach president", what does it mean?
  - Increase in interest or salience of the issue (information seeking)?
  - Or, increase in support for impeaching the president (expressing preference)?
- Generally, it is difficult to infer attitudes from behavior based on observations alone
  - Behaviors are often overdetermined: different causes led to same observed behavior
  - This is a broad problem we often face beyond the big data context

Limitation 4: Causal Inference

- Finding interesting patterns (covariation) is itself a daunting task, because a hallmark of Big Data is the fact that it vastly exceeds human comprehension (Shiffrin, 2016).

- Additionally, there are enormous difficulties facing researchers trying to draw causal inference from or about some pattern found in Big Data.
  - There are almost always a large number of additional and mostly uncontrolled confounders and covariates (Z) with correlations among them
  - This is particularly the case given that most Big Data are formed as a nonrandom sample taken from the infinitely complex real world.

Limitation 5: Data Quality

- Researchers do not control how the data is generated and data quality
- (In)Consistency over time
  - Data is consistently produced but algorithms that generate the data could be changed time to time (e.g. updating system and adjusting platforms and services)
- (Lack of) Veracity
  - Much of information from the internet is fake
  - Algorithms can be gamed: astroturfing