Big Data

Nathan TeBlunthuis (University of Washington)

Abstract

This article is to orient practicing social scientists who might work with "Big Data." What is Big Data and what is it good for? What are its limitations? How does one use it effectively? What ethical concerns does it raise?

Big Data allows social and political researchers to study great quantities of observations of people and their communications and behaviors. Yet what precisely constitutes Big Data is slippery and has changed over time as new data technologies and practices have spread. Here, the term describes the use of big datasets for social or political research. This is primarily traces of behaviors and messages on social media but can also include data collected through sensors such as GPS devices and mobile phones, or government records. Generally, Big Data has so many observations or measured variables that it requires special computational, statistical, and ethical consideration.

Keywords: big data, quantitative methods, computational social science, social media

Defining Big Data

Although many social and political researchers are interested in working with Big Data, the term "Big Data" is a buzzword used in marketing products like databases, statistical packages and "artificial intelligence" (another marketing buzzword). As trends in data technologies have evolved, so have the types of datasets and technologies marked with Big Data. Therefore, Big Data often describes data-work of a scale requiring tools or techniques that are not yet commonplace. In many fields and organizations, spreadsheets have been the dominant tool for data analysis and Big Data might refer to a dataset too large to work with in a spreadsheet, but that can be analyzed using a statistical programming language like R or Python on a typical laptop. As tools like R have gained popularity among social scientists, the meaning of Big Data may evolve to require new specialized computing infrastructure and skills, like high-performance computers or map-reduce systems.

In recent years, advanced open-source systems have rapidly made Big Data work much more accessible. Software developed under the Apache project is particularly powerful and mature. Apache Spark is a map-reduce framework that can be used in multiple programming languages or through a SQL-style syntax, process datasets that are too large to fit in memory and distribute computation across many computers on academic HPC systems. Apache Parquet stores data in a columnar format which can make reading subsets of the data much more efficient and is enhanced by Apache Arrow, which can efficiently select, filter, and modify data as it is read to save memory.

Social scientists do not always need to use such technologies to study Big Data because they can often obtain samples which they can analyze with familiar tools. Even still, they should be aware of the power and pitfalls of the Big Data from which these samples are drawn.

Large samples and statistical significance

The obvious appeal of Big Data for social research is that large sample sizes have enormous statistical power. Traditional study designs like surveys or laboratory experiments can be underpowered and fail to rule out small but theoretically interesting relationships. Although a large sample size is not the only benefit that comes from Big Data, it is the least ambiguous one. Yet just because a correlation can be measured does not make it important. Big Data analyses can detect extremely weak correlations that might not reflect or theoretically significant relationships. The statistical training and practice of social and behavioral scientists often places the greatest emphasis on "hypothesis tests" that compare an observed statistic with a "null hypothesis" of no measurable relationship or correlation. Instead of emphasizing a null hypothesis of no relationship, an effective Big Data study will consider what types of results a reasonable scholar should consider minimally interesting and interpret statistical results in substantive terms.

For example, consider Bond et. Al. (2012)'s famous experiment on 6.3 million Facebook users in which they randomly showed potential voters messages designed to prompt voting behavior. The University of California San Diego publicized the study in a press release with the headline "Facebook Boosts Voter Turnout,"[1] but how much did Facebook's intervention really matter? The researchers detected that the messages increased voting behavior with a p-value of 0.02 and helpfully interpret their effect size as showing that the intervention increased the probability of voting by 0.39%. It is conceivable that an intervention like this could tip the scales in an extremely close election. Yet the effect is so small that it would have been extremely difficult to detect without a Big Data experiment run by a major social media platform.

That said, the statistical power of Big Data comes with the responsibility to understand what the variables included in the data represent and how the data are limited. Early excitement about the potential of Big Data to revolutionize social science has given way to a more sober understanding that big datasets also have important limitations. Big datasets are rarely constructed for research purposes and therefore may be missing important variables, unrepresentative of general populations, and have other issues that can threaten the validity of research.

Particularly important is that background knowledge is often vital to understanding results from big data studies. For example (Tsvekova et al., 2017) found that found that bots often undo each other's Wikipedia edits and concluded that "even good bots fight." However, (Geiger and Halfaker, 2017) argue that these bots were not "fighting" but undoing each other's edits by design. Tsvetkova et al.'s study was methodologically rigorous, but Geiger and Halfaker's deeper expertise in Wikipedia bots was important to correctly interpreting their findings. Robust scientific conclusions depend not only on Big Data and methodological sophistication but also on accumulated qualitative knowledge that can ground interpretations.

Reducing Bias in Big Data

[1] https://ucsdnews.ucsd.edu/pressrelease/facebook_fuels_the_friend_vote

Even with good contextual knowledge, Big Data can mislead. Many fields of social science, economics most of all, highly value analyses that afford causal inference (Morgan and Winship, 2007; Pearl 2009). Unlike the Facebook voter turnout experiment, most large datasets provide only observational data and are not constructed with experimental controls. Therefore correlations within them cannot easily be given a causal interpretation. Big Data does not automatically make causal inference easier, rather it can be particularly vulnerable to spurious and misleading findings (Calude and Longo, 2017).

Efforts to develop robust methodologies for drawing causal inferences from observational (non-experimental) data are of increasing interest to researchers in social science, statistics, and computer science (Morgan and Winship, 2007; Pearl 2009). Attempts to draw causal inferences from non-experimental data depend on assumptions that are fundamentally untestable or that may be known to be wrong. Big Data, especially from social media and other digital traces, are often incomplete in ways that limit statistical adjustment. Still, reducing bias in statistical estimation can improve the quality of evidence for causal theories that big data can provide.

A common source of bias in Big Data is that data can be missing in unknown ways (Tufekci, 2014). Collecting, transmitting, storing, and analyzing large datasets requires expensive computational resources. Researchers often work with samples of data to reduce these costs. However, if not done carefully, non-representative sampling can easily introduce bias. Sampling bias can be only be fully corrected using statistical adjustment if the probability of each data point in sample and in the population is known. This is often difficult if not impossible using social media data (Morgan and Winship, 2007). However, APIs like the Twitter Search API (V1) can return incomplete data without giving any information about which data points were included or removed making it impossible to know if a sample is representative and providing no way to correct the bias if not.

Statistical adjustment is also important for reducing bias by accounting for the influence of confounding variables. For example, including demographic variables like age, sex and race a regression model helps explain the correlation between musical tastes and political attitudes (DellaPosta, 2015). Demographic variables are very often related to individual attitudes and behaviors, but they are very often missing in Big Data from social media (Salganik, 2018).

Where statistical adjustment in regression models attempts to reduce bias in parameter estimates by adjusting for confounding variables, a quasi-experimental research design uses information about how the treatment was non-randomly assigned to support causal inference. In a true experiment, the investigator randomly assigns a *treatment* and can therefore assume that a correlation between the treatment and the outcome represents a causal effect. Quasi-experimental research designs attempt to approximate the analysis of a controlled experiment using observational data and can sometimes work well with Big Data.

When a treatment is assigned non-randomly such that some individuals or groups are more likely to receive the treatment than others, an analyst can attempt to adjust for the treatment assignment process using *matching methods.* The idea is to create a "balanced" dataset where each member of the treatment group is compared to a highly similar member of the control

group. This way, even if some factors make individuals more likely to be treated than others, an analyst can draw fair comparison in the balanced dataset where members of the treatment and control groups have these factors in equal proportion. For example, Lelkes (2016) used constructed a balanced dataset of over 100,000 survey responses to provide evidence that Europeans whose party loses elections that exposed to greater amounts of political news are more likely to lose faith in institutions and satisfaction with Democracy.

There are many different strategies for constructing a "balanced" dataset. The historically most popular strategy, called propensity score matching, has recently been shown to be brittle and inefficient and alternative strategies such as coarsened exact matching are increasingly popular (King and Nielsen, 2019). Big Data can be advantageous for matching because chances are good that comparable members of the treatment and control group can be found. However, an important limitation is that one can only match based on observed variables. It is not possible to use matching to correct for unobserved factors that make some individuals more likely to receive the treatment than others, which are often missing in Big Data.

Powerful quasi-experiments can be conducted using Big Data sources that are "always on" (Salganik, 2018), collecting data continuously over time. For example, changes to social media algorithms for news feeds or recommendations may influence behavior. If not accounted for, such changes can introduce bias, but researchers can use knowledge of how the algorithms work, or the timing of design changes for causal inference. For example, TeBlunthuis et al. (2021) study an algorithm for predicting misbehavior on Wikipedia using a regression discontinuity design (RDD), a type of quasi-experiment that uses a *discontinuity* in the relationship between a *forcing variable* that determines treatment assignment and the outcome to set up a fair comparison between treatment and control groups (Morgan and Winship, 2007). In TeBlunthuis et al. (2021), the discontinuity is when a score output by the algorithm crosses an arbitrary threshold that triggers the treatment, which was being flagged in a moderation tool. The validity of an RDD depends on using a linear regression to statistically adjust for the relationship between the forcing variable and the outcome. When sample sizes are as large as the number of edits to Wikipedia, an RDD analysis can analyze only data points very close to the threshold and thereby require less statistical adjustment.

With knowledge of the timing of a sudden change, researchers can use an interrupted time series analysis to draw a before-and-after comparison in a comparable way to an RDD. Analyses of data at only one point in time can be biased by fixed attributes of subjects, but with repeated observations over time a within-subjects panel data analysis is robust to all time-invariant confounders (Morgan and Winship, 2007). Keep in mind that panel data analyses can still be biased by time-varying confounders like unknown changes in algorithms or platform designs. Without sufficient variation in predictors and outcomes, a panel data analyses can be underpowered, so Big Datas' large sample sizes can be a great boon.

Ethics of Big Data Collection and Analysis

Big datasets can raise distinctive ethical concerns that are essential to understand and consider before collecting, analyzing, or publishing big datasets (boyd and Crawford, 2012; Zimmer, 2018). Big Data research in particular risks violating individuals' expectations of privacy and

how their information may be used and has even involved experimentation on individuals without their informed consent. Conventional institutional safeguards like Institutional Review Boards (in a US context) may not recognize all the potential threats of Big Data or even see the analysis of public social media posts as "human subjects research" requiring their oversight. This places additional responsibility on researchers to protect individuals whose data are collected and analyzed from potential harm.

For example, in 2016, Emil O. W. Kirkegaard, collected and released data from over 70,000 user profiles on the dating site OKCupid to widespread criticism (Zimmer, 2018). Even though these profiles were technically available to anyone with an OKCupid account, Kirkegaard's dataset made these profiles much more accessible than they were before. Zimmer (2018) draws from the ethicist Helen Nissenbaum to argue that what Kierkegaard did was wrong because it breached the *contextual integrity* of individuals' information in ways that exposed them to new types of harms. OKCupid users intended to share profiles and personality questions with potential social and romantic partners who were also using OKCupid, not with the public.

The systems that generate and collect Big Data are often designed for commercial and administrative purposes that may violate norms of research ethics. Social media platforms and search engines regularly experiment on their users while developing and testing features. These experiments are not necessarily innocuous. When researchers at Cornell University and Facebook reported on their experimental manipulation of the emotional content of Facebook user's feeds, the public reaction was swift and negative (Hallinan et al., 2019). Facebook users did not consent to be experimented on, but the Cornell IRB (Institutional Review Board) ruled that because Facebook conducted the study was except from oversight. Social media companies are not held to the same legal standards as academic researchers, yet their experimentation can still violate social norms against treating people as "Guinea Pigs" by experimenting on them without their consent.

Understanding the distinction between ethics and legality is essential. Social media companies have ubiquitous "Terms of Service" (TOS) for legal protection yet violating TOS can be ethical in some circumstances (Fiesler et al., 2020). For example, the Ad Observer research project at NYU collected data from volunteer study participants who installed a browser extension and gave their informed consent to participate in the study investigating how Facebook targets political advertisements (Vincent, 2021). The NYU researchers seemed to be doing important and ethical research. Yet Facebook blocked their research saying their activities violated TOS. Kirkegaard's actions were unethical not because they TOS in scraping OKCupid data, but they violated users' expectations and exposed them to new harms. On the other hand, if TOS, such as Twitter's, say that data may be used in research, this should not be understood as establishing informed consent because few users even read TOS (Hallinan et al., 2020).

Data from settings where expectations of privacy are low raise fewer ethical risks. For example, Wikipedia is a setting where expectations of privacy because inspecting the historical activity of Wikipedia editors is an understood and widespread practice that helps ensure that the encyclopedia is accurate and accountable. Even in settings where individuals

may not expect data on their communications or activities to be analyzed in research, they may be more comfortable with analyses of a vast number of individuals than research that scrutinizes a few (Hallinan et al., 2020). This has been a brief and complete overview of the ethical concerns Big Data can raise. Research ethics will continue to evolve as researchers, publics and institutions gain experience with the collection and analysis of Big Data.

References

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

boyd, danah, & Crawford, K. (2012). Critical Questions For Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Calude, C. S., & Longo, G. (2017). The Deluge of Spurious Correlations in Big Data. *Foundations of Science*, *22*(3), 595–612. https://doi.org/10.1007/s10699-016-9489-4

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why Do Liberals Drink Lattes? *American Journal of Sociology*, *120*(5), 1473–1511. https://doi.org/10.1086/681254

Fiesler, C., Beard, N., & Keegan, B. C. (2020). No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 187–196.

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, *4*(1), 2056305118763366. https://doi.org/10.1177/2056305118763366

Geiger, R. S., & Halfaker, A. (2017). Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of "Even Good Bots Fight." *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 49:1-49:33. https://doi.org/10.1145/3134684

Hallinan, B., Brubaker, J. R., & Fiesler, C. (2020). Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society*, *22*(6), 1076–1094. https://doi.org/10.1177/1461444819876944

King, G., & Nielsen, R. (2019). Why Propensity Scores Should Not Be Used for Matching. *Political Analysis*, *27*(4), 435–454. https://doi.org/10.1017/pan.2019.11

Lelkes, Y. (2016). Winners, Losers, and the Press: The Relationship Between Political Parallelism and the Legitimacy Gap. *Political Communication*, *33*(4), 523–543. https://doi.org/10.1080/10584609.2015.1117031

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*.

TeBlunthuis, N., Hill, B. M., & Halfaker, A. (2021). Effects of Algorithmic Flagging on Fairness: Quasi-experimental Evidence from Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 56:1-56:27. https://doi.org/10.1145/3449130

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of Wikipedia. *PLOS ONE*, *12*(2), e0171774. https://doi.org/10.1371/journal.pone.0171774

Tufekci, Z. (2014, May 16). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *Eighth International AAAI Conference on Weblogs and Social Media*. Eighth International AAAI Conference on Weblogs and Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062

Vincent, J. (2021, August 4). *Facebook bans academics who researched ad transparency and misinformation on Facebook*. The Verge. https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in

Zimmer, M. (2018). Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media + Society*, *4*(2), 2056305118768300. https://doi.org/10.1177/2056305118768300

Further Recommended Literature (if needed)

boyd, danah, & Crawford, K. (2012). Critical Questions For Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*.