ANNUAL REVIEWS

Annual Review of Sociology

Data ex Machina: Introduction to Big Data

David Lazer[1,2] and Jason Radford[1,3]

[1]Department of Political Science and College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115; email: d.lazer@neu.edu, j.radford@neu.edu

[2]Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts 02138

[3]Department of Sociology, University of Chicago, Chicago, Illinois 60637

Annu. Rev. Sociol. 2017. 43:19–39

First published as a Review in Advance on May 17, 2017

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-060116-053457

Keywords

big data, computational social science, technology, research methodology, quantitative sociology, research ethics, research design, social media, CDR data, data linkage, networks, mobility

Abstract

Social life increasingly occurs in digital environments and continues to be mediated by digital systems. Big data represents the data being generated by the digitization of social life, which we break down into three domains: digital life, digital traces, and digitalized life. We argue that there is enormous potential in using big data to study a variety of phenomena that remain difficult to observe. However, there are some recurring vulnerabilities that should be addressed. We also outline the role institutions must play in clarifying the ethical rules of the road. Finally, we conclude by pointing to a number of nascent but important trends in the use of big data.

INTRODUCTION

Archives of human activity go back millennia; however, the increasingly comprehensive digital archives of human behavior, combined with the exponential growth of computational power, create the potential for a transformation of fields such as sociology. Core constructs of sociology, such as interaction, collective action, expression, and diffusion of behavior lurk in these archives. It is possible to study the connectivity of entire human societies, including who communicates with whom about what, how people move through space, who says what, and who buys what, all with a temporal granularity of seconds. The coming generation will witness a transformation of sociological theory through these improvements in our ability to observe dynamic social systems (boyd & Crawford 2012, Golder & Macy 2014).

Big data does pose distinctive challenges for scholars. These digital archives are not the product of scientific design. The information captured in these archives is not what a social scientist would choose. Further, what is captured is constantly, and sometimes abruptly, changing. The signal in big data is vulnerable to manipulation, sometimes purposive, sometimes incidental. Further, relevant behaviors are split into many archives, with no practical way of conjoining them. For example, there may not be a strong conceptual or empirical distinction or boundary to be drawn between behaviors captured by different cell phone carriers, but most research on cell phone data is based on data from a single carrier.

Big data also presents enormous institutional challenges to sociology. The large majority of sociologically relevant analysis of big data is done by computer scientists, and there is relatively little reflection of the big data revolution in top sociology journals. Only 6 of 182 articles published between 2012 and 2016 in the *American Journal of Sociology* (*AJS*) and 9 of 240 in *American Sociological Review* (*ASR*) involve the use of big data.[1]

The objective of this review, therefore, is to critically assess the potential of big data for the collective intellectual endeavor that is sociology. In what follows, we provide a brief overview of what big data is and then review recent big data projects. We break these projects down into three types of big data, enumerate the promises and pitfalls common across them, and offer guidance to sociology for moving forward. We conclude by drawing together the challenges that remain across all three areas and outline the work that needs to be done to put the field on more solid ethical, methodological, and epistemic ground. Our hope is to demonstrate the value of big data research while articulating its most pressing problems and offering reasonable strategies for advancing the field.

WHAT IS BIG DATA?

The term "big data" has proliferated across society and owes its recent popularization to a report by McKinsey (Manyika et al. 2011). The issue of scale of data is relevant across the academy, from astronomy to the humanities. The manifestation of big data, of course, varies. In astronomy, big data takes the form of images that take petabytes of storage; in the humanities, it can take the form of millions of digitized books.

The dominant data paradigm for the quantitative social sciences has been tabular—variables in columns, cases in rows. Big sociological data, in contrast, comes in an infinite array of other

[1]For our review, we looked for any studies using big data sources or existing data with big data methods such as automated text analysis. We especially looked for social media data, data from online platforms like crowdfunding and online dating websites, data from smartphones including CDR data, and large scale administrative data sets. At *AJS*, these works include Burt (2012), Knigge et al. (2014a), Legewie (2016), Legewie & Schaeffer (2016), Lin & Lundquist (2013), and de Vaan et al. (2015). For *ASR*, the relevant literature includes Bail (2012), Curington et al. (2015), Diekmann et al. (2014), Goldberg et al. (2016), Hall et al. (2015), Knigge et al. (2014b), Leung (2014), van de Rijt et al. (2013), and Vasi et al. (2015).

forms, such as pictures, video, tweets, text, space, networks, streaming, etc. In early formulations, its "bigness" meant that the data could not be processed with extant software (Gartner 2011, Manovich 2012). Big data was big to the extent that new technologies had to be created by specialists to collect, store, and analyze it. The term big data thus refers to data that are so large (volume), complex (variety), and/or variable (velocity) that the tools required to understand them must first be invented (Laney 2001, Monroe 2013). Big data thus requires a computational social science—a mash-up of computer science for inventing new tools to deal with complex data, and social science, because the substantive substrate of the data is the collective behavior of humans (Lazer et al. 2009).

The big data challenges of today have analogues in history. Two generations ago, running multivariate linear models required punch cards and a computer the size of a basement. The challenge was to build tools that leveraged ever-increasing processing power. Today, the technical and theoretical challenges confronting big data research are different, stemming from the complex and heterogeneous form of contemporary big data. These challenges, too, can be met. Sociologists must learn how to cull through large amounts of unstructured text data, program mobile phones, and build data pipelines that scrape, process, store, and make available large amounts of data of wide varieties and types.

Big Data Sources

Big data is vast and heterogeneous, encompassing everything from YouTube to digital archives of books. A comprehensive review of all big data sources is beyond the scope of this review. There are many discrete literatures around different big data sources, and even a complete list of those literatures would soon be obsolete. Instead, we offer a fuzzy typology of big data sources, based on the loci of data collection. We begin with a discussion of digital life—the capturing of digitally mediated social behaviors—that likely accounts for the majority of big data research. We then discuss digital trace data, the archival exhaust of the modern bureaucratic organization, and conclude with digitalized life data, the movement of intrinsically analog behavior into digital form.

Digital life. An increasing fraction of life is intrinsically digitally mediated. Twitter, Facebook, and Wikipedia are all platforms where behaviors are all online. These behaviors may relate to offline occurrences, but behaviors like tweeting are inherently digital. Behaviors on such platforms are typically substantially captured by platform owners, because their business models rely on inferences (such as ad targeting) that can be made from these data streams. Further, it is possible for third parties to harvest data from these platforms. Facebook allows users to download portions of their data; the entire edit history of Wikipedia can be downloaded for analysis; Google allows some access to search query volume. Twitter data are the most used by scholars because of their accessibility. Third parties can negotiate and pay for access for large samples of Twitter data or harvest sizable targeted subsamples.

Digital platform data may be viewed in two ways. The first is to view these platforms as generalizable microcosms of society (Tufekci 2014). Kossinets & Watts (2006) examine email to study the role of different social foci in emergent network structure. Barberá et al. (2015) use Twitter to study political mobilization. The dissemination of news and rumors has been studied on Wikipedia (Keegan et al. 2013, Keegan & Brubaker 2015), Twitter (Bakshy et al. 2011, Romero et al. 2011, Yang & Counts 2010), and across platforms (Goel et al. 2012, Kim et al. 2014). Online markets such as Airbnb and Kickstarter have been used to study patterns of social inequality (Edelman & Luca 2014, Greenberg & Mollick 2017). The Billion Prices Project uses price data scraped from online retailers to track inflation (Cavallo & Rigobon 2016). In this vein too is

Google Flu Trends, which sought to track cases of flu using Google search data (Ginsberg et al. 2009).

The second way to view these platforms is as distinctive realms in which much of the human experience now resides. Exemplary in this regard is the research on whether Facebook creates or accentuates an informational filter around individuals such that people only see ideologically compatible content (Bakshy et al. 2015; also see Lazer 2015). The objective of this article is not to resolve the filter bubble question more generally—the study's finding of only moderate filtering effects tells us nothing about Google or Twitter, for example. And its entire scientific relevance rests on the (correct) assumption that Facebook is, by itself, general enough to be worthy of study.

We note that not all behavior on a platform may be captured. Indeed, the entire Internet may be viewed as an enormous platform for human expression and, although there are efforts to capture records of the Internet, most notably by the Internet Archive, generally these capture only snapshots of particular moments for particular URLs.

Further, as we discuss below, the mapping between what happens on these platforms and the phenomena of interest to a sociologist may be somewhat weak. All friends are not Facebook friends, and not all Facebook friends are friends.

Digital traces. The modern complex organization creates a steady output of records that chronicle actions taken (sometimes labeled metadata). Call detail records (CDRs) from phone calls are illustrative (Onnela et al. 2007, Toole et al. 2015). CDRs from cell phones, for example, typically offer time stamps and duration of calls, identifiers for initiator and recipient of calls, and identifiers for the cell towers accessed during the call. Such information has been used to critically examine the strength of weak ties (Onnela et al. 2007), predict individual and collective level unemployment (Toole et al. 2015), and model the spread of malaria (Wesolowski et al. 2012). Governmental data, such as voter records, political contribution data (Bonica 2014), and tax data (Chetty et al. 2016) are other examples. What distinguishes trace data from digital life is that the trace is only a record of the action, not the action itself.

Digitalized life. Lastly, digitalized life represents the capture of nonintrinsically digital life (i.e., most of life) in digital form. Thus, for example, phones can be programmed to continuously identify nearby Bluetooth devices, thus capturing the proximity of individuals (Eagle et al. 2009). The constant video recording of major modern cities creates ongoing records of human interaction. Informational objects that predate computers can be easily scanned into manipulable digital form. Exemplary in this regard are Google Books and related endeavors (Michel et al. 2011), which involved scanning millions of books, as well as the use of newspaper data to study the dynamics of fame (van de Rijt et al. 2013).

Instrumentation of human behavior. Crosscutting these three types of big data is the possibility for the proactive and purposive instrumentation of human behavior. It is possible to identify and collect data on subsets of behaviors: to monitor a stream of certain types of tweets, track web browsing behaviors from particular locations, or scrape select data from select websites, such as with the Billion Prices Project (Cavallo & Rigobon 2016). Smartphones can be instrumented to collect a wide array of ambient data (e.g., Eagle et al. 2009). Specially designed hardware, so-called sociometers, can track minute details regarding face-to-face social interactions, and have been used, for example, to study the role of gender in collaboration (Onnela et al. 2014).

These examples highlight the variety of uses for these new data. These projects rely heavily on a new breed of tools and techniques, largely from computer science.

Summary

Big data represents both the new kinds of digital data available to sociologists as well as the tools and technologies required to access these data. The explosion of applications of big data to long-standing social questions reveals a variety of important opportunities for sociologists to extend our knowledge. In the following, we outline these opportunities and provide suggestions for taking advantage of them. These applications, too, both in their successes and their failures, reveal new challenges for big data research, which must be addressed if we are to advance in a responsible, equitable, and scientifically sound way.

OPPORTUNITIES

Our review of the literature reveals a set of distinct opportunities present in a world replete with big data. Much of these data are what is generally called massive, passive data: data generated in the process of meaningful social behavior rather than data reported for research. The prevalence of such systems and digital devices means that whole systems are captured by these data. And because these systems are always running, they act as controls for experiments offline, online, and even in the lab. One final opportunity in big data comes from making big data small by finding the special populations within it (Foucault Welles 2014).

Massive, Passive: Behavioral Data at Scale

In principle, big data archives offer measures of actual behaviors, as compared with self-reports of behaviors. The literature is rife with evidence of the problems with self-reported behavior. Generally, self-reported behavior is noisy, with a variety of systemic biases. For example, people systematically lie about everything from whether they voted, to what their weight and height are. Certain types of behavior are entirely inaccessible via self-reports.

To focus on one example, social network researchers have long struggled to accurately measure social ties (Bernard et al. 1984, Marsden 1990). Respondents are biased in the ties they remember and how they remember them. Respondents provide very different networks when asked who they go to for advice, who they would ask for a loan, who their friends are, and who they spend the most time with. And the same question may be interpreted differently by different people. These differences present a range of problems for survey-based social network research, including ambiguity in defining what a tie means, difficulty interpreting concepts such as social influence, and uncertainty about omitted ties.

Behavioral data enable us to observe social networks through interactions. Eagle et al. (2009) compare self-report data to behavioral data. They handed out 94 smartphones to students for nine months. The phones were programmed with applications that "recorded and sent the researcher data about call logs, Bluetooth devices in proximity of approximately five meters, cell tower IDs, application usage, and phone status" (2009, p. 15274). The researchers constructed a behavioral measure of participants' social networks using the call log data and spatial proximity captured via Bluetooth. Participants were behaviorally related if they talked to one another or were in physical proximity. In the middle of the study, participants filled out a questionnaire about their physical and social proximity to one another. The survey constituted the self-report measure of students' networks.

Comparing reported physical proximity to behavioral measures of proximity, Eagle et al. find recency and salience biases in subjects' recall. People tended to overreport physical proximity in general. Additionally, friends were much more accurate in recalling their physical proximity to

one another than those who did not consider one another friends. The researchers also find that subjects who reported being friends behaved in stable and substantially different ways from those who reported not being friends. Friends were much more likely to be spatially proximate to one another at night or on the weekends even if they were at work or somewhere else. Moreover, self-reported friends in January exemplified this pattern five months later in May.

This study helps validate passive network inference by showing that behavioral measures can capture self-reported friendship. At the same time, it reveals a substantial shortcoming of self-reported measures in capturing our weak ties—the colleagues and acquaintances who we see on a regular basis but who we do not consider close friends.

The Copenhagen Network Study (Stopczynski et al. 2014a,b) extends behavioral research on social networks by comparing different modes of behavioral inference. Researchers handed out 1,000 phones to students entering the Technical University of Denmark in 2012 and 2013. Researchers used the phones to infer Bluetooth proximity, geographical proximity via GPS, and interaction via calls and text messages. They combined this with students' Facebook data, their proximity to routers, and qualitative field observations from an anthropologist. They find that nearly the entire call network is captured by Bluetooth proximity, and 80% of students' Facebook friends could be captured by Bluetooth proximity. Yet only 20% of students' Facebook friend network was captured by call logs. Stopczynski et al. conclude that passive behavioral measures are not interchangeable and different digital systems can lead to different social networks with different properties. That is, there is not a single social network for anyone, but a series of shifting networks based on the organizations and technologies individuals use to form and sustain relationships. Thus, the choice of which systems to use to collect data from subjects and how to integrate these different data sources will likely affect study results.

Nowcasting

Monitoring social phenomena is an essential part of social science research. Surveys such as the University of Michigan's Survey of Consumers or the Bureau of Labor Statistics' Current Employment Statistics survey are fielded regularly to monitor the economic health of the United States. The Centers for Disease Control and Prevention and World Health Organization use a network of testing labs, healthcare providers, and government agencies to generate regular estimates of flu prevalence and virulence.

These regularly updated, vital statistics monitor phenomena that are essential to the workings of contemporary institutions. However, these surveillance systems are very expensive to operate, time consuming to deploy, and inaccurate at high levels of temporal and geographic granularity. Through the digitization of social life, these phenomena are increasingly becoming visible in big data. Such digitization offers the potential to reduce the costs, improve the accuracy, and increase the scale of societal monitoring.

Researchers have used existing large-scale digital data such as CDR data and scraped web data as sensors for monitoring social phenomena. CDR data have been used to detect unemployment (Toole et al. 2015). Google Flu Trends sought to track cases of flu using Google search data (Ginsberg et al. 2009; although see Lazer et al. 2014). Finally, Beauchamp (2016) and Hopkins & King (2010) use Twitter data to generate estimates of public opinion. Each of these studies represents an attempt to validate the use of big data to generate estimates of socially relevant phenomena. This area of research, called nowcasting, seeks to generate descriptions of the world as they happen.

The Billion Prices Project is one such ambitious project. Using price data scraped from online retailers, Cavallo, Rigobon, and colleagues produce daily estimates of prices that are essential to

everything from exchange rates to inflation to the real value of wages (see Cavallo & Rigobon 2016 for an overview). Much like traditional price indexing, researchers at the Billion Prices Project scrape data for a preselected "bag" of goods from a curated set of online retailers from around the world. They compare the prices among the same goods within retailers within countries every day to compute a daily consumer price index for over 70 countries.

This approach has unearthed several new empirical insights. The standard data used to study price changes showed that price changes were normally distributed, with most changes being small. However, Cavallo (2017) finds that the distribution of price changes is actually bimodal, with large price increases and decreases, and very few small price changes. Cavallo argues that the price imputations made during the construction of standard data produce small but erroneous price changes. In essence, measurement error led to a decade's worth of poor theory on how prices change.

A second result of the Billion Prices Project has been to document violations of the law of one price (Cavallo et al. 2014). This law asserts that a good should generally have the same price no matter where it is sold, controlling for differences in the value of currency. Cavallo et al. show that the law of one price only holds among countries that share a currency.

Finally, Cavallo (2013) demonstrates the social impact of nowcasting. For example, the government of Argentina began manipulating its estimates of inflation in 2007. Cavallo attempts to identify the extent of manipulation by estimating the divergence between Argentina's official estimates and those derived from online prices. As a control, Cavallo also compares Consumer Price Indexes computed via online prices to those published by the governments of Brazil, Venezuela, Chile, and Colombia. The results show that although indexes computed using online prices track prices in the control countries, they diverge substantially in the case of Argentina. In 2015, Argentina stopped publishing inflation numbers, and the Billion Prices Project has been used to infer inflation in their place.

The Billion Prices Project demonstrates some of the implications of using big data to estimate social phenomena. First, it acts as an alternative method of estimating social phenomena, illuminating problems in traditional methods. Second, it can act as a measure of something that is otherwise unmeasured or whose measure may be disputed. Finally, nowcasting demonstrates big data's potential to generate measures of more phenomena, with better geographic and temporal granularity, much more cost-effectively than traditional methods. It is worth noting that this granularity is most useful when fused with traditional methods rather than used as a replacement for them (Lazer et al. 2014).

Data on Social Systems

Perhaps most exciting about big data is the opportunity to build a science of society, a science that would study society at scale, composed of subsystems and individuals that are dynamically connected in particular ways and locations. For example, State et al. (2015) examine networks based on interactions among millions of people on Twitter and between Yahoo! email users. Both data sets revealed strong intracultural correlations among individuals across the globe.

Analysis of the aggregate activity of Twitter reveals expected and unexpected patterns in global behavior. For example, there are strong diurnal patterns across the globe (Golder & Macy 2011). People are generally happier in the morning and earlier in the week, but these patterns vary substantially with work and season and differ systematically for so-called night owls. Dodds et al. (2011) show that big data can capture annual emotional cycles, particularly around holidays, as well as divergences during global events such as the financial crisis.

Data on systems have been used to answer long-standing questions about human mobility. Online data and data from mobile phones have been used to characterize human mobility (Brockmann

et al. 2006, González et al. 2008). This mobility information has been linked to geographical, cultural, and political information in order to study patterns of interaction at the national and regional levels (Blanford et al. 2015, Sevtsuk & Ratti 2010).

One such study is that of Toomet et al. (2015), who use cell phone data to examine the patterns of interaction among members of the Estonian majority and Russian-speaking minority members in the Estonian capital of Tallinn. They used CDR data from Estonia's largest mobile provider, Telia Eesti (formerly EMT), to capture individuals' locations to within several hundred meters. When individuals made calls or sent texts in a shared location at a particular time, Toomet et al. counted that as copresence. To distinguish Estonian and Russian-speaking users, they used the preferred language settings that cell phone providers link to a phone's SIM (subscriber identity module) card. Finally, Toomet and colleagues inferred individuals' home and work locations based on their most frequent geographic location during work and evening hours.

Toomet et al. analyze the ethnic dissimilarity of copresent mobile users at work, home, and during free time. They find high similarity (i.e., substantial segregation) among individuals at home and at work. However, they find substantial dissimilarity among individuals during free time periods. This mixing during free time was most prominent in the city's central district, but also existed even in the less integrated suburbs. The findings suggest that although major social institutions engender segregation, people desegregate in less structured social life. The study raises questions for the long history of research on segregation and social inequality, which almost uniformly define segregation as residential segregation (Massey & Denton 1993, Wilson 1987). As Small (2004, ch. 5) suggests, segregation is as much a process of where you live and work as it is the cultural and physical boundaries that shape who can and cannot participate in the life of a neighborhood. This study shows that high-resolution social mobility data offer a more nuanced picture of segregation bridging residence, employment, and community.

Natural and Field Experiments

Digitally mediated social systems such as Facebook and Twitter and newly digital institutions like the Internal Revenue Service (IRS) capture data irrespective of events in the environment. In addition, the administrators of these systems make innumerable changes to them over time. As such, it is plausible that all kinds of natural experiments may be hidden within large-scale data. Big data offers a milieu for studying the effect of external events on ongoing social processes. For example, Phan & Airoldi (2015) use Facebook data to examine how social networks were affected by Hurricane Ike in 2008. Ayers et al. (2011) use search traffic data to determine whether tax increases on cigarettes lead to an increased interest in tobacco cessation.

Big data can capture the effects of experiments in the field through data linkage. Exemplary in this regard is a series of studies by Chetty and collaborators who merge IRS data with data from prior research (Chetty et al. 2014a,b, 2016). Chetty et al. (2016) linked participants in the Moving to Opportunity (MTO) field experiment to their tax filings with the IRS decades later. MTO sought to investigate neighborhood effects by examining the impact of receiving vouchers and moving to wealthier and safer neighborhoods on the economic, social, and psychological well-being of low-income families. By linking IRS data to the MTO data, Chetty et al. (2016) were able to measure the effect of the experimental treatment on children's future earnings. They found that the economic impact only occurred for children who were younger when they moved, indicating that neighborhood effects require long exposure periods to affect children's future earnings.

Big data systems themselves can create natural experiments by changing user behavior through subtle and not-so-subtle changes to their policies and practices. For example, Brown et al. (2010) use changes to eBay's interface to test the effect of suppressing price information, specifically shipping costs, on purchase decisions. Researchers can use changes to policies, designs, or algorithms

as experimental manipulations. The very malleability of digital worlds make them powerful vessels for conducting very large experiments.

Facebook has been the setting for several large-scale field experiments using randomized manipulations of what people could see about their peers to study social influence (Bond et al. 2012, Kramer et al. 2014). For example, Bond et al. conducted a field experiment using all Facebook users in the United States who were over the age of 18 and who logged into Facebook on the date of the US election in 2012. Six hundred thousand users were put into an informational condition and received a message encouraging them to vote, providing information on where to vote, and allowing them to click an "I Voted" button. Sixty million users were randomly put into the social condition and received the same information but with a list of all of their friends who clicked the "I Voted" button. Finally, a control group of six hundred thousand users saw no message at all.

They examined the extent to which study participants actually voted, engaged with the "I Voted" button, or clicked the link to view voter information. To measure actual voting they linked a subsample of the study population to public voting records. They found a direct effect: Those in the informational condition were more likely to vote than those in the control group, and those in the social condition were more likely than those in the informational condition to vote. They also found an indirect effect: Those who had friends in the social condition were 0.255% more likely to vote per friend in the condition and 0.012% more likely to seek out information per friend in the condition. These findings are important, proving that online social networks contribute to the diffusion of offline behavior. It presents another insight beyond this. The effects of the study are miniscule but, in aggregate, still represent hundreds of thousands of voters. Big data systems offer an unprecedented degree of precision in measuring small but meaningful effects.

Making Big Data Small

The power of big data is often the small data contained within. Although many online platforms such as Reddit, Wikipedia, and Microsoft's Xbox Live are dominated by young, white, Western men, within their millions of users, one can still find hundreds or thousands of people who are older, nonwhite, female, and non-Western. Studies that "make big data small" (Foucault Welles 2014) either use big data to observe traditionally hard to reach populations or utilize the vast array of very specific kinds of cases to generate robust estimates.

Big data provides access to data on traditionally underrepresented populations. Twitter data have been used to study people who suffer from PTSD, suicidal ideation, and depression (Coppersmith et al. 2014; De Choudhury et al. 2013, 2016). Jackson & Foucault Welles (2016) and Barberá et al. (2015) use Twitter to identify individuals at the center of emergent social movements including Black Lives Matter, the Gezi Park protest in Turkey, Occupy Wall Street, and the Spanish Indignados.

Barberá et al. use Twitter to compare the patterns of mobilization between a successful protest mobilization at Taksim Gezi Park in 2013 (one locus of the Arab Spring) and two unsuccessful mobilizations by Occupy Wall Street and the Indignados in the spring of 2012. They searched Twitter's public application program interface (API) for keywords and hashtag keywords to collect samples of tweets from the three movements. They also collected two other samples of tweets from widespread, nonprotest activity to act as a control (the 2014 Academy Awards and a year's worth of tweets related to raising the minimum wage in the United States). They constructed mobilization networks among users who posted messages containing the keywords and hashtags and users who reposted those messages (i.e., retweets).

They find peripheral members of the Gezi Park protest mobilized more people than the core Occupy and Indignados protesters did. They find no such core-periphery patterns in the nonpolitical mobilization cases. They argue that successful mobilization involves getting peripheral

members to recruit still more peripheral people to protest: The mobilized must mobilize others themselves. This adds to contemporary social movement theory, showing that resources and organizational capacity are not enough to mobilize protest. Network diffusion can separate success from failure.

As studies of online protest show, using big data for research takes advantage of the increasing digitization of social life. Social movement studies have largely depended on newspapers and organizational archives because they are the few sources regularly logging social movement activities (Earl et al. 2004). Social media have become an essential mode for people at the margins of society to connect with one another and express thoughts and sentiments that otherwise go unsaid.

A second use for the small data within big data is in the robust estimates they can generate from the many narrow samples contained within. One example of this is the use of big data to generate population estimates. Paramount in this vein is Wang et al.'s (2015) use of surveys from a panel of Xbox users to predict national election polls in the United States. Xbox users skew heavily toward young, white, Western men. However, even though men made up 93% of the panel, the poll of roughly 340,000 still contained roughly 24,000 women. Using multilevel regression and poststratification, Wang et al. generate estimates of state and national polls for "demographic cells" (Wang et al. 2015, p. 981) such as college educated women over 55 in Florida who identify as Republican and voted for McCain in 2008. They aggregate these cell estimates to predict polls at the state and national level. They find that this reweighting of highly skewed big data accurately predicts national and most state-level polls for the presidential race and Obama's eventual election victory.

In addition, large samples contain enough unusual cases to robustly estimate heterogenous effects. Small data sets are blunt tools able only to detect large average effects. However, many associations of interest in sociology are contingent on individual and contextual factors. The study of intersectionality is premised on the belief that the main forces in society differ by the particular combinations of race, class, gender, sexuality, and other identities (Collins 1998). Big data allows for robustly estimating effects at the intersection. Others are developing algorithms that seek to inductively identify groups for whom effects are especially large (Athey & Imbens 2016, Green & Kern 2012, Imai & Ratkovic 2013, Taddy et al. 2016).

As survey researchers have long known, a well-defined sample of even a small size can tell us more than millions of poorly defined cases (Squire 1988). These studies by Wang et al., Athey & Imbens, and others show that we can recover the power of well-defined, small data from big data. The difference is that big data and big populations like Xbox users already exist, making them substantially more cost efficient to use to gather data than traditional sample building methods.

VULNERABILITIES

The core issue with any data is who and what get represented. With surveys, one might ask, for example, which respondents are accessible, and what they can accurately reveal. The scale of big data sets creates the illusion that they contain all relevant information on all relevant people. However, the difference between big and everything is still infinite, and the core issues of social science research around validity and generalizability still apply. Further, certain big data can be quite brittle, vulnerable to changes in the data generation process and to attacks motivated by the fact that they are materially consequential.

Generalizability

Big data are almost always convenience samples offering a distinct set of advantages and disadvantages. However, the data now easily available are unlike most convenience samples hitherto

common in the social sciences. Many are often convenience censuses: a complete record of a certain set of individuals or behaviors that match certain criteria. Scale and seeming comprehensiveness of data often obscure major issues regarding inclusion and selection and therefore representativeness and generalizability.

Many big data census efforts aspire to capture all possible data but do so without a systematic sampling frame. For example, projects like EventRegistry (Leban et al. 2014) and GDELT (Leetaru & Schrodt 2013) identify global events from major media sources such as the *New York Times* and Associated Press, news aggregators such as Google News and LexisNexis, and regional news sources, which are added continuously. They crawl as many sources as possible with the goal of creating as near a census of world events as possible. Yet these projects lack principles for sampling on critical features such as geography, publication frequency, journalistic practices, and type of news. This means that different kinds of events like sporting events or weather may be covered in geographically and temporally uneven ways. Beyond this, more data cannot solve the long-standing problems of using news sources to study events (Earl et al. 2004).

"Big data hubris" (Lazer et al. 2014, p. 1203) is the belief that volume can solve all problems. With near-census projects, trying to create a census without a sampling frame causes errors associated with selection, missing data, and thin coverage to be inestimable (Japec et al. 2015). To use an analogy, these near-census projects are like sending millions of people out into the streets to count the population of the United States. You would count a large number of people, but you could not know the kinds of people who are counted twice or not counted at all, and therefore you could not know what kinds of people are over- or underrepresented and to what extent.

However, big data often is a census of a particular, conveniently accessible social world. All of Twitter is a census of Twitter. Data from Kickstarter are a census of Kickstarter. However, even census data have limits when they come from a single platform. Tufekci (2014) notes that Twitter has become to social media scholars what the fruit fly is to biologists—a model organism. Tufekci argues that relying on a single platform produces issues for generalizability unique to model organisms. For example, different social media platforms have different rules for following or friending others and posting or sending messages. Thus, the patterns of friendship and diffusion differ across platforms (Tufekci 2014, p. 507). Different model organisms behave differently and therefore can give divergent results. In addition, platforms differ in the age, gender, race, and class backgrounds of their users (Perrin 2015). Even phone call data can exclude or overrepresent certain populations (Pestre et al. 2016). Not all social processes and forces are well-represented in a given model organism.

Generalizability is always a question of reference: To what do we want to generalize? With big data, the widespread availability of convenient census and near-census data leads many to overstate the reach of their findings. As research using multiple social media platforms shows, the results from one population of users do not necessarily apply to another. And convenient near-census data may play to our big data hubris that volume will trump sampling. The solution is to use data from multiple sources to validate the findings from any one of them. This will take a commitment from scientists to create institutional structures that can provide access to these data sources. In addition, near-census projects need to take sampling more seriously in order to estimate error in the data produced (Japec et al. 2015).

Too Many Big Data

Being tied to individual platforms presents another problem: when the scientifically relevant behavior spans these platforms. For example, consider trying to measure a simple construct like "who an individual regularly talks with." People use different modes to interact with different people,

such as text messaging, face-to-face, or by phone. And people can use functionally equivalent tools within a mode, such as cell phones, land lines, Skype, and Google Hangouts to reach different people. Thus, the data on who someone interacts with exist in a variety of different data sets. As society continues to develop new ways for people to interact with one another, our data on social interaction will continue to fragment. We call this the problem of too many big data.

This issue is even worse for many important sociocultural constructs, such as friendship, which are arguably more cognitive and normative than behavioral. How does one observe love, affection, or deceit from cell phone data?

These issues with too many big data are potentially surmountable. They point to the need to create a fusion of emerging computational methods with existing social science methods. For example, one study has shown promise in tracking interaction across digital and physical modes. As we discussed earlier, the Copenhagen Network Study has proven we can track interaction across digital and physical modes. The results provide a rare glimpse into the ways in which these data can overlap or diverge from one another. However, this multimode data collection is expensive and intrusive.

Other work could compare use of a particular platform relative to population averages to understand the potential limits of the platform. It is also plausible that nonbehavioral constructs like friendship, love, and trust are observable in the sense that there may be certain behaviors strongly correlated with certain cognitive constructs. To identify the best signals, we need a Rosetta stone connecting behavioral big data constructs to theoretically motivated social constructs (Margolin et al. 2013).

Artifacts and Reactivity

The caveats to big data extend beyond the problems of only studying a specific data set. Big data systems are themselves susceptible to various kinds of error and misappropriation. In the next section, we discuss how social forces can manipulate these data in unexpected and difficult-to-detect ways. In this section, we talk about artifacts, the errors and anomalies that systems produce, and reactivity, the changes in data resulting from technical changes rather than underlying changes in behavior.

Platforms do not merely represent data but generate them. In some cases, it can be difficult to distinguish observations resulting from errors in the system from those representing a real change in underlying behavior. For example, in the Google Ngram project, the word "fuck" is used with startling frequency in books published through 1800, and drops to near zero during the 1800s. Upon closer inspection, it is clear that this did not reflect some dramatic shift in social mores, but rather is an artifact of contemporary optical character recognition systematically misinterpreting an archaic version of "s" as an "f."[2]

When platforms change how they operate, both behavior and the way behavior is recorded can change. Earlier we argued that these changes can act as natural experiments. Here, we show these changes can have negative consequences for science. Google Flu Trends was perhaps one of the most prominent uses of big data, estimating the prevalence of flu using Google search traffic (Ginsberg et al. 2009, Ortiz et al. 2011). However, it seems likely that Google changed its search to make it more useful for finding health-related information, leading people to perform more searches for the flu during the peak of the season. The result was that the number of

[2] **http://searchengineland.com/when-ocr-goes-bad-googles-ngram-viewer-the-f-word-59181**.

searches increased relative to the number of flu cases, and Google Flu Trends started dramatically overreporting the flu (Lazer et al. 2014).

Behavior on any platform is part emergent, part mechanical. The hashtag convention on Twitter was not imagined or proposed by the company but instead originated from users who found keywords searchable when they began with the hash. Only later did Twitter make these hashtags into hyperlinks. There will always be changes in a platform, and artifacts inevitably slip in. Social scientists must be aware of and wary of the technical and social history of the platforms from which their data are generated.

The Ideal User Assumption: Bots, Puppets, and Manipulation

In big data analysis, we often assume that the data have been generated by a specific type of user, most often single, unique people who express themselves honestly through their personal accounts. This ideal user assumption fails to hold under a wide variety of critical circumstances. Many accounts are not operated by human beings. Additionally, users can have multiple accounts, sometimes with the intent to conceal the user's true identity. Finally, people, organizations, and even nation states put platforms to unintended uses. Taken together, the characteristics of the ideal user need to be validated rather than assumed, and nonideal users generating big data need to be studied in their own right.

The first type of violation of the ideal user assumption is when we incorrectly assume all users are human. This is most often violated in the case of robots and organizations. It is likely that robots are present in all big data systems, whether bots on social media (Ferrara et al. 2016), nonplayer characters in online games (Lee & Ramler 2015), or even robocalls in CDR data (Gupta et al. 2015). Many organizations and institutions are present and active in these systems as well, including corporations, governments, and terrorist organizations (McCorriston et al. 2015). Robots and organizations engage on these platforms for a variety of prosocial and antisocial purposes, making them difficult to detect and theorize (Lee et al. 2011, Ferrara 2015).

The ideal-user assumption is also violated when we assume everyone is who they claim to be. Humans misrepresent themselves in a variety of ways outlined by Wang et al. (2006). Individuals can create multiple accounts for different purposes, including alters to express different identities or so-called sock puppets to fabricate support for themselves or their cause (Bu et al. 2013, Zheng et al. 2006). Individuals can also be deceptive about critical aspects of their identity, as in the case of catfishing. Finally, people can use a fraudulent identity, presenting themselves to be another particular person.

The final violation of the ideal user assumption is when users manipulate the platform in unintended ways to achieve surreptitious goals. In analyzing big data, we often assume that the data are being generated by people behaving in good faith. Yet, many users attempt to game the system or create rigged systems (Ferrara 2015). News media have reported cases in which countries and election campaigns have used robots and real people on social media to artificially inflate their own positions and influence (Durgin 2016, Matthew 2015). In 2016, the Chinese peer-to-peer lending platform Ezubao was revealed to be a Ponzi scheme in which 95% of the proposals were fake (Xinhua 2016). Finally, actors, particularly nation states, can directly control these platforms. As King et al. (2014) find, the Chinese government conducts mass censorship of online media to suppress news of any events with the potential to yield extragovernmental political mobilizations such as protests.

Social media platforms that are highly permeable and very cheap to manipulate are easy targets for robots, deception, and manipulation. However, even high cost, impermeable platforms like CDR and newswire data can be subject to this activity. The challenge is that each of these violations

of the ideal user assumption typically occurs under different conditions. Sock puppetry and political manipulation are probably more likely to occur in contentious arenas and topics, whereas fraud and market manipulation are probably likely to occur at the intersection of the financially vulnerable and the financially influential. In analyzing big data, sample control is critical for making inferences about human beings acting in good faith. However, it is important to remember that these forms of nonideal-user behavior are worthy of study on their own.

RESEARCH ETHICS

There are major ethical issues regarding the acquisition and use of big data for researchers, institutions, and society at large. Some issues are new, but many are new versions of long-standing issues. The problem, however, is that there is no consensus on what the rules should be, and the policies and recommendations set forth by scientific associations vary substantially, often contradicting one another. Rules will eventually become clear, but the risks to researchers, universities, and the public remain high until they do. Further, these ethical issues in turn suggest interesting researchable questions, ranging from issues around reidentification (Sweeney 2002) to the meaning and management of consent by subjects (Stopczynski et al. 2014a,b).

The National Research Council proposed to amend the definition of human subjects research to "a systematic investigation designed to develop or contribute to generalizable knowledge by obtaining data about a living individual directly through interaction or intervention, or by obtaining identifiable private information about an individual" (NRC 2014, recommendation 2.1, p. 40). However, rules regarding obtaining subject consent are typically obviated when the data have been collected by a third party and have been anonymized. Is this an acceptable standard given the prevalence of collaboration between companies and academics? Furthermore, given the ease of deidentifying many kinds of data, what standards can be set for anonymization?

Regulating informed consent, the heart of human subjects research, is only one of the central issues that have yet to be settled. Other open issues include the rights of secondary subjects, measuring harms from the loss of privacy, and regulating the status of leaked data like the Panama Papers, which revealed international tax evasion by the world's elite, or the data dump of the US adultery website Ashley Madison. The role of the university becomes critical here because it is part of the regulatory apparatus enforcing rules and protecting scholars, but also because it provides the training infrastructure that empowers compliance.

FUTURE TRENDS

Researchers have used big data to answer old questions in new ways and new questions never before answerable. Their successes and failures have helped us identify the promises and pitfalls of research in this area and the kinds of investments institutions need to make for the future of this research. In this final section, we point to six trends that will likely affect the big data landscape in hopes of helping sociology get ahead of the curve.

More Data Are Coming

Big data will continue to grow into more domains. For example, EventRegistry provides data on events, and it also acts as a repository providing real-time access to more than 100,000 news publishers (**http://www.EventRegistry.org**). Big data will also continue to reach back into the past as libraries digitize their collections, newspapers digitize their archives, and initiatives such as

Google Books and Project Gutenberg digitize books. The question for data coverage will continue to be less about whether the data exist, but more about what can be studied with the available data.

More linkages between different big data will become more common. The work of Chetty and colleagues (2014a,b, 2016) with IRS data is only the beginning. Another opportunity is connecting online data to offline data, such as linking social media accounts to voter records and data collected by brokers such as Acxiom. One revolutionary form of big data linkage is wikification, which involves linking words and phrases in text to entities in Wikipedia (Mihalcea & Csomai 2007). Entity linkage allows researchers to use the structured and unstructured data Wikipedia maintains on entities to enhance the contextual information associated with texts.

Different Data Are Coming

The majority of data being created on big data platforms are still unusable for social scientific research. These are the images, audio, and video being created, discussed, and shared. The tools for providing meaningful structure to these data at scale have lagged behind those of other types of data, such as text. They are quickly catching up. For example, image processing using convolutional neural nets and other deep learning methods has become as easy to use in Python and R as methods for text analysis. Furthermore, the tools to analyze these data are increasingly being made available through publicly accessible interfaces like Google Cloud Vision API. With publicly accessible models, researchers upload their files to the service, which uses pretrained models to make inferences about the file and then sends these inferences back as metadata.

Models Will Become More Generic

Google Cloud Vision API is a prime example of another critical trend in machine learning: creating generic models and making them available to the public. There is a long precedent of creating and sharing models for processing unstructured data. The Linguistic Inquiry and Word Count dictionary is perhaps the best known in the social sciences (Tausczik & Pennebaker 2010). However, such models are now being published in a variety of methods. For example, Jozefowicz et al. (2016) trained a deep learning model on the One Billion Word Benchmark and are publishing the model itself as an alternative to the model released in 2014 in Stanford's GloVe (Pennington et al. 2014). Like the Vision API, researchers can use these models on their own texts to generate word embeddings.

Generic models allow researchers to use pretrained machine learning models on their own data, rather than having to deal with the issues of data processing and model specification. These out-of-the-box machine learning projects aspire to use big data to create the most effective models and then make those models the standard for processing unstructured data. However, generic models are not necessarily better at applied tasks than specialized models. And, in the absence of social theory, such generic models may miss obvious social patterns in data, potentially reinforcing long-standing social biases (Caliskan et al. 2017).

Data from Multiple Platforms Will Become Standard

As big data systems proliferate and multiple systems offer similar services, it will become increasingly possible and easier for researchers to perform studies on different platforms. The CrowdBerkeley project is offering data for multiple crowdfunding websites such as Kickstarter, Indiegogo, and Kiva. Multiple forms of big data will also be used to create parallel measures. For example, new models of political partisanship are being generated by Federal Election

Commission data, Twitter, press releases, and floor speeches (Barberá 2015, Bonica 2014, Gentzkow et al. 2016, Tsur et al. 2015). Interestingly, each of these measures provides different narratives about the emergence of partisanship and demonstrates the importance of using different data to approach the same phenomena.

Qualitative Approaches to Big Data

While big data is typically viewed as quantitative social science on steroids, we anticipate innovative approaches weaving together qualitative methods and computational approaches to large-scale data. There is a long history of archival research in the social sciences. Digital archives present the challenge of vast amounts of information beyond the capacity of armies of grad students to read. Searching and sorting of archives becomes essential to qualitative understanding. At the simplest level, this might simply require keyword searches, but certainly more complex, computationally enabled approaches will emerge. For example, consider hypothetical research examining the Internet Archive's version of **http://www.congress.gov**. The snapshots of the members' home pages present too large a data set to comprehensively read, but it is feasible to query the data for policy statements on health care by every member for targeted reading and hand-coding.

Methodological Integration

The prior point highlights a more general lesson: Big data will increasingly be integrated with existing research methods in sociology. Big data offers strengths and weaknesses that are quite different than existing data sources (Lazer et al. 2014). The most compelling sociological research in the twenty-first century will not be big data but a fusion of data sources related to important questions. Survey data will be linked to a tiny portion of archival data, providing inferential power to the entire archive that it otherwise would not have. Interesting or typical cases in big data can be identified for qualitative exploration. The scientific payoff should, in turn, be insight into phenomena that heretofore have been neglected, related to the connectivity and dynamics of entire societies.

The future of big data is as bright and fraught as its past. While sociology has generally lagged behind in using big data, there are many opportunities for the field to take advantage of and many challenges and debates to confront. Further, the increasing presence of digitally mediated social activity and the increasingly digital social life mean that the need to integrate big data approaches into sociology will increase for the foreseeable future, with the corresponding need for sociologists to contribute to our understanding of an increasingly digital and digitalized world.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF0920053 (the ARL Network Science CTA) and in part by a grant from the US Army Research Office W911NF1210556. Any opinions expressed are the authors' alone.

LITERATURE CITED

Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113(27):53–60

Ayers JW, Ribisl K, Brownstein JS. 2011. Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 "SCHIP" cigarette tax increase. *PLOS ONE* 6(3):e16777

Bail CA. 2012. The fringe effect. *Am. Sociol. Rev.* 77(6):55–79

Bakshy E, Hofman JM, Mason WA, Watts DJ. 2011. Everyone's an influencer: quantifying influence on twitter. *Proc. 4th ACM Conf. Web Search Data Mining*, pp. 65–74. New York: ACM

Bakshy E, Messing S, Adamic LA. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239):1130–32

Barberá P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Polit. Anal.* 23(1):76–91

Barberá P, Wang N, Bonneau R, Jost JT, Nagler J, et al. 2015. The critical periphery in the growth of social protests. *PLOS ONE* 10(11):1–15

Beauchamp N. 2016. Predicting and interpolating state-level polls using Twitter textual data. *Am. J. Political Sci.* 61:490–503

Bernard HR, Killworth P, Kronenfeld D, Sailer L. 1984. The problem of informant accuracy: the validity of retrospective data. *Annu. Rev. Anthropol.* 13:495–517

Blanford JI, Huang Z, Savelyev A, MacEachren AM. 2015. Geo-located tweets. Enhancing mobility maps and capturing cross-border movement. *PLOS ONE* 10(6):e0129202

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, et al. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–98

Bonica A. 2014. Mapping the ideological marketplace. *Am. J. Polit. Sci.* 58(2):367–86

boyd d, Crawford K. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* 15(5):662–79

Brockmann D, Hufnagel L, Geisel T. 2006. The scaling laws of human travel. *Nature* 439(7075):462–65

Brown J, Hossain T, Morgan J. 2010. Shrouded attributes and information suppression: evidence from the field. *Q. J. Econ.* 125(2):859–76

Bu Z, Xia Z, Wang J. 2013. A sock puppet detection algorithm on virtual spaces. *Knowl.-Based Syst.* 37:366–77

Burt RS. 2012. Network-related personality and the agency question: multirole evidence from a virtual world. *Am. J. Sociol.* 118(3):543–91

Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334):183–86

Cavallo A. 2013. Online and official price indexes: measuring Argentina's inflation. *J. Monet. Econ.* 60(2):152–165

Cavallo A. 2017. Scraped data and sticky prices. *Rev. Econ. Stat.* In press. **http://dx.doi.org/10.1162/REST_a_00652**

Cavallo A, Neiman B, Rigobon R. 2014. Currency unions, product introductions, and the real exchange rate. *Q. J. Econ.* 129(2):529–95

Cavallo A, Rigobon R. 2016. The Billion Prices Project: using online prices for measurement and research. *J. Econ. Perspect.* 30(2):151–78

Chetty R, Friedman JN, Rockoff JE. 2014a. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* 104(9):2593–632

Chetty R, Friedman JN, Rockoff JE. 2014b. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* 104(9):2633–79

Chetty R, Hendren N, Katz LF. 2016. The effects of exposure to better neighborhoods on children: new evidence from the Moving to Opportunity experiment. *Am. Econ. Rev.* 106(4):855–902

Collins PH. 1998. It's all in the family: intersections of gender, race, and nation. *Hypatia* 13(3):62–82

Coppersmith G, Harman C, Dredze M. 2014. Measuring post traumatic stress disorder in Twitter. *Proc. 8th Int. AAAI Conf. Weblogs Soc. Media*, pp. 579–82. **http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8079**

Curington CV, Lin K-H, Lundquist JH. 2015. Positioning multiraciality in cyberspace. *Am. Sociol. Rev.* 80(4):764–88

De Choudhury M, Gamon M, Counts S, Horvitz E. 2013. Predicting depression via social media. *Proc. 7th Int. AAAI Conf. Weblogs Soc. Media*, pp. 128–37. **http://course.duruofei.com/wp-content/uploads/2015/05/Choudhury_Predicting-Depression-via-Social-Media_ICWSM13.pdf**

De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Proc. 2016 CHI Conf. Hum. Factors Comput. Syst.*, pp. 2098–2110. New York: ACM Press

De Vaan M, Vedres B, Stark D. 2015. Game changer: the topology of creativity. *Am. J. Sociol.* 120(4):1144–94

Diekmann A, Jann B, Przepiorka W, Wehrli S. 2014. Reputation formation and the evolution of cooperation in anonymous online markets. *Am. Sociol. Rev.* 79(1):65–85

Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM. 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLOS ONE* 6(12):e26752

Durgin C. 2016. Inside Donald Trump's Potemkin Twitter army. *National Review*, Apr. 8. **http://www.nationalreview.com/article/433870/donald-trumps-twitter-supporters-might-be-fake**

Eagle N, Pentland AS, Lazer D. 2009. Inferring friendship network structure by using mobile phone data. *PNAS* 106(36):15274–78

Earl J, Martin A, McCarthy JD, Soule SA. 2004. The use of newspaper data in the study of collective action. *Annu. Rev. Sociol.* 30(1):65–80

Edelman BG, Luca M. 2014. *Digital discrimination: the case of Airbnb.com*. Working Pap. 14–054, NOM Unit, Harvard Bus. Sch.

Ferrara E. 2015. "Manipulation and abuse on social media" by Emilio Ferrara with Ching-man Au Yeung as coordinator. *ACM SIGWEB Newsletter*, *Spring* 4:1–9

Ferrara E, Varol O, Davis C, Menczer F, Flammini A. 2016. The rise of social bots. *Commun. ACM.* 59(7):96–104

Foucault Welles B. 2014. On minorities and outliers: The case for making big data small. *Big Data Soc.* 1(1):1–2

Gartner. 2011. *Gartner says solving "big data" challenge involves more than just managing volumes of data*. News Release, June 27. **http://www.gartner.com/newsroom/id/1731916**

Gentzkow M, Shapiro JM, Taddy M. 2016. Measuring polarization in high-dimensional data: method and application to congressional speech. NBER Work. Pap. 22423, Natl. Bur. Econ. Res., Cambridge, MA

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–14

Goel S, Watts DJ, Goldstein DG. 2012. The structure of online diffusion networks. *Proc. 13th ACM Conf. Electron. Commer.*, pp. 623–38. New York: ACM

Goldberg A, Srivastava SB, Manian VG, Monroe W, Potts C. 2016. Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. *Am. Sociol. Rev.* 81(6):1190–222

Golder SA, Macy MW. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–81

Golder SA, Macy MW. 2014. Digital footprints: opportunities and challenges for online social research. *Annu. Rev. Sociol.* 40:129–52

González MC, Hidalgo CA, Barabási A-L. 2008. Understanding individual human mobility patterns. *Nature* 453(7196):779–82

Green DP, Kern HL. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* 76(3):491–511

Greenberg J, Mollick E. 2017. Activist choice homophily and the crowdfunding of female founders. *Adm. Sci. Q.* 62:341–74

Gupta P, Srinivasan B, Balasubramaniyan V, Ahamad M. 2015. *Phoneypot: data-driven understanding of telephony threats*. Brief. Pap., NDSS Symp. 2015, San Diego, CA

Hall M, Crowder K, Spring A. 2015. Neighborhood foreclosures, racial/ethnic transitions, and residential segregation. *Am. Sociol. Rev.* 80(3):526–49

Hopkins DJ, King G. 2010. A method of automated nonparametric content analysis for social science. *Am. J. Political Sci.* 54(1):229–47

Imai K, Ratkovic M. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* 7(1):443–70

Jackson SJ, Foucault Welles B. 2016. #Ferguson is everywhere: initiators in emerging counterpublic networks. *Inf. Commun. Soc.* 19(3):397–418

Japec L, Kreuter F, Berg M, Biemer P, Decker P, et al. 2015. *AAPOR report on big data*. Am. Assoc. Public Opin. Res., Oakbrook Terrace, IL

Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y. 2016. Exploring the limits of language modeling. arXiv:1602.02410 [cs.CL]

Keegan BC, Brubaker JR. 2015. "Is" to "was": coordination and commemoration in posthumous activity on Wikipedia biographies. *Proc. 18th ACM Conf. Comput. Support. Coop. Work Soc. Comput.*, pp. 533–46. New York: ACM

Keegan BC, Gergle D, Contractor N. 2013. Hot off the wiki: structures and dynamics of Wikipedia's coverage of breaking news events. *Am. Behav. Sci.* 57(5):595–622

Kim M, Newth D, Christen P. 2014. Trends of news diffusion in social media based on crowd phenomena. *Proc. 23rd Int. Conf. World Wide Web*, pp. 753–58. New York: ACM

King G, Pan J, Roberts ME. 2014. Reverse-engineering censorship in China: randomized experimentation and participant observation. *Science* 345(6199):1–10

Knigge A, Maas I, van Leeuwen MHD. 2014a. Sources of sibling (dis)similarity: total family impact on status variation in the Netherlands in the nineteenth century. *Am. J. Sociol.* 120(3):908–48

Knigge A, Maas I, van Leeuwen MHD, Mandemakers K. 2014b. Status attainment of siblings during modernization. *Am. Sociol. Rev.* 79(3):549–74

Kossinets G, Watts DJ. 2006. Empirical analysis of an evolving social network. *Science* 311(5757):88–90

Kramer ADI, Guillory JE, Hancock JT. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111(29):8788–90

Laney D. 2001. *3D data management: controlling data volume, velocity and variety*. Res. Note, META Group, Stamford, CT. **https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf**

Lazer D. 2015. The rise of the social algorithm. *Science* 348(6239):1090–91

Lazer D, Kennedy R, King G, Vespignani A. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343(6176):1203–5

Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, et al. 2009. Life in the network: the coming age of computational social science. *Science* 323(5915):721

Leban G, Fortuna B, Brank J, Grobelnik M. 2014. Event Registry: learning about world events from news. *Proc. 23rd Int. Conf. World Wide Web*, pp. 107–10. New York: ACM

Lee C-S, Ramler I. 2015. Rise of the bots: bot prevalence and its impact on match outcomes in League of Legends. *Int. Worksh. Netw. Syst. Support Games (NetGames), Zagreb, Dec. 3–4*, pp. 1–6

Lee K, Eoff BD, Caverlee J. 2011. Seven months with the devils: a long-term study of content polluters on Twitter. *5th Int. AAAI Conf. Weblogs Soc. Media.* **https://pdfs.semanticscholar.org/1dd5/355e62b9fc37a355e135d5909ed28128d653.pdf**

Leetaru K, Schrodt PA. 2013. GDELT: global data on events, location, and tone, 1979–2012. *Int. Stud. Assoc. Annu. Conf., San Diego*. **http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.686.6605&rep=rep1&type=pdf**

Legewie J. 2016. Racial profiling and use of force in police stops: how local events trigger periods of increased discrimination. *Am. J. Sociol.* 122(2):379–424

Legewie J, Schaeffer M. 2016. Contested boundaries: explaining where ethnoracial diversity provokes neighborhood conflict. *Am. J. Sociol.* 122(1):125–61

Leung MD. 2014. Dilettante or Renaissance person? How the order of job experiences affects hiring in an external labor market. *Am. Sociol. Rev.* 79(1):136–58

Lin K-H, Lundquist J. 2013. Mate selection in cyberspace: the intersection of race, gender, and education. *Am. J. Sociol.* 119(1):183–215

Manovich L. 2012. Trending: the promises and the challenges of big social data. In *Debates in the Digital Humanities*, Vol. 2, ed. MK Gold, pp. 460–75. Minneapolis, MN: Univ. Minn. Press

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, et al. 2011. *Big data: the next frontier for innovation, competition, and productivity*. Rep., McKinsey Global Inst. **http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation**

Margolin D, Lin Y-R, Brewer D, Lazer D. 2013. Matching data and interpretation: towards a Rosetta stone joining behavioral and survey data. *7th Int. AAAI Conf. Weblogs Soc. Media*, pp. 9–10. **http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6267**

Marsden PV. 1990. Network data and measurement. *Annu. Rev. Sociol.* 16:435–63

Massey DS, Denton NA. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard Univ. Press

Matthew S. 2015. Revealed: how Russia's "troll factory" runs thousands of fake Twitter and Facebook accounts to flood social media with pro-Putin propaganda. *The Daily Mail*, March 28

McCorriston J, Jurgens D, Ruths D. 2015. Organizations are users too: characterizing and detecting the presence of organizations on Twitter. *9th Int. AAAI Conf. Web Soc. Media*. **http://www-cs.stanford.edu/~jurgens/docs/mccorriston-jurgens-ruths_icwsm-2015.pdf**

Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–82

Mihalcea R, Csomai A. 2007. Wikify!: Linking documents to encyclopedic knowledge. *Proc. 16th ACM Conf. Inf. Knowl. Manag.*, pp. 233–42. New York: ACM

Monroe BL. 2013. The five Vs of big data political science: introduction to the Virtual Issue on Big Data in Political Science. *Polit. Anal.* 19(5):66–86

NRC (Natl. Res. Counc.). 2014. *Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences*. Washington, DC: Natl. Acad. Press

Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, et al. 2007. Structure and tie strengths in mobile communication networks. *PNAS* 104(18):7332–36

Onnela J-P, Waber BN, Pentland A, Schnorf S, Lazer D. 2014. Using sociometers to quantify social interaction patterns. *Sci. Rep.* 4:5604

Ortiz JR, Zhou H, Shay DK, Neuzil KM, Fowlkes AL, Goss CH. 2011. Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. *PLOS ONE* 6(4):1–9

Pennington J, Socher R, Manning CD. 2014. GloVe: global vectors for word representation. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pp. 1532–43

Perrin A. 2015. *Social networking usage: 2005–2015*. Rep., Pew Res. Cent., Washington, DC

Pestre G, Letouzé E, Zagheni E. 2016. *The ABCDE of big data: assessing biases in call-detail records for development estimates*. Presented at Annu. Bank Conf. Dev. Econ., June 20–21, Washington, DC. **http://pubdocs.worldbank.org/pubdocs/publicdoc/2016/6/551311466182785065/Pestre-Letouze-Zagheni-ABCDE-May-2016.pdf**

Phan TQ, Airoldi EM. 2015. A natural experiment of social network formation and dynamics. *PNAS* 112(21):6595–600

Romero DM, Meeder B, Kleinberg J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. *Proc. 20th Int. Conf. World Wide Web*, pp. 695–704. New York: ACM

Sevtsuk A, Ratti C. 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *J. Urban Technol.* 17(1):41–60

Small ML. 2004. *Villa Victoria: The Transformation of Social Capital in a Boston Barrio*. Chicago: Univ. Chicago Press

Squire P. 1988. Why the 1936 *Literary Digest* poll failed. *Public Opin. Q.* 52(1):125–33

State B, Park P, Weber I, Macy M. 2015. The mesh of civilizations in the global network of digital communication. *PLOS ONE* 10(5):e0122543

Stopczynski A, Pietri R, Pentland A, Lazer D, Lehmann S. 2014a. Privacy in sensor-driven human data collection: a guide for practitioners. arXiv:1403.5299 [cs.CY]

Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, et al. 2014b. Measuring large-scale social networks with high resolution. *PLOS ONE* 9(4):e95978

Sweeney L. 2002. K-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 10(05):557–70

Taddy M, Gardner M, Chen L, Draper D. 2016. A nonparametric Bayesian analysis of heterogenous treatment effects in digital experimentation. *J. Bus. Econ. Stat.* 34(4):661–72

Tausczik YR, Pennebaker JW. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29(1):24–54

Toole JL, Lin Y-R, Muehlegger E, Shoag D, González MC, Lazer D. 2015. Tracking employment shocks using mobile phone data. *J. R. Soc. Interface* 12(107):20150185

Toomet O, Silm S, Saluveer E, Ahas R, Tammaru T. 2015. Where do ethno-linguistic groups meet? How copresence during free-time is related to copresence at home and at work. *PLOS ONE*. 10(5):e0126093

Tsur O, Calacci D, Lazer D. 2015. A frame of mind: using statistical models for detection of framing and agenda setting campaigns. *Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Int. Joint Conf. Nat. Lang. Process.*, *Beijing, July 26–31*, pp. 1629–38. **https://pdfs.semanticscholar.org/f5c8/dbeea0112227486b7fc3bd20a73726ffea88.pdf**

Tufekci Z. 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. arXiv:1403.7400 [cs.SI]

van de Rijt A, Shor E, Ward C, Skiena S. 2013. Only 15 minutes? The social stratification of fame in printed media. *Am. Sociol. Rev.* 78(2):266–89

Vasi IB, Walker ET, Johnson JS, Tan HF. 2015. "No fracking way!" Documentary film, discursive opportunity, and local opposition against hydraulic fracturing in the United States, 2010 to 2013. *Am. Sociol. Rev.* 80(5):934–59

Wang GA, Chen H, Xu JJ, Atabakhsh H. 2006. Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* 36(5):988–99

Wang W, Rothschild D, Goel S, Gelman A. 2015. Forecasting elections with non-representative polls. *Int. J. Forecast.* 31(3):980–91

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. 2012. Quantifying the impact of human mobility on malaria. *Science*. 338(6104):267–70

Wilson WJ. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: Univ. Chicago Press

Xinhua. 2016. Online P2P lender suspected of $US 7.6 billion fraud. *Xinhua*, Feb. 1. **http://news.xinhuanet.com/english/2016-02/01/c_135065022.htm**

Yang J, Counts S. 2010. Predicting the speed, scale, and range of information diffusion in Twitter. *ICWSM* 10:355–58

Zheng R, Li J, Chen H, Huang Z. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.* 57(3):378–93