JOURNAL OF INFORMATION TECHNOLOGY & POLITICS
2017, VOL. 14, NO. 2, 97–109
https://doi.org/10.1080/19331681.2017.1312187

Routledge
Taylor & Francis Group

GUEST EDITORIAL

The empiricist's challenge: Asking meaningful questions in political science in the age of big data

Andreas Jungherr and Yannis Theocharis

ABSTRACT
The continuously growing use of digital services has provided social scientists with an expanding reservoir of data, potentially holding valuable insights into human behavior and social systems. This has often been associated with the terms "big data" and "computational social science." Using such data, social scientists have argued, will enable us to better understand social, political, and economic life. Yet this new data type comes not only with promises but with challenges as well. These include developing standards for data collection, preparation, analysis, and reporting; establishing more systematic links between established theories within the existing body of research in the social sciences; and moving away from proofs-of-concepts toward the systematic development and testing of hypotheses. In this article, we map these promises and challenges in detail and introduce five highly innovative contributions collected in this special issue. These articles illustrate impressively the potential of digital trace data in the social science all the while remaining conscious of its pitfalls.

KEYWORDS
Big data; computational social science; digital trace data; methodology

Digital trace data in the social sciences: Promises and challenges

The continuously growing use of digital services has provided social scientists with an expanding reservoir of data, potentially holding valuable insights into human behavior and social systems. The potentials of the use of digital trace data in social science research has famously given rise to the terms "big data" and "computational social science." Using such data, social scientists have argued, will enable us to better understand social, political, and economic life through the generation of large data sets composed not of answers to questions asked of citizens concerning their attitudes and behaviors, but of the digital traces documenting their actual behavior as they use digital devices and services.

Although the potential of the use of digital trace data has been a continuous focus in public debate, scientific contributions using these data in political science usually come in the form of research manifestos or isolated proofs-of-concepts, only marginally contributing to current debates in the social sciences. Currently, most work using digital trace data in the analysis of political phenomena falls into two categories. In the first category fall studies using digital trace data to illustrate online components of political events, such as protests, televised debates, or election campaigns. The second category collects studies demonstrating that in specific cases, specific selections of digital trace data collected on specific services somewhat resemble routinely used metrics in political science. Here, authors tend to conclude that digital trace data allow the identification and prediction of political phenomena.

Even though there are many interesting and valuable contributions among studies using digital trace data, to move into the mainstream of political science research the field has to mature. This includes: developing standards for data collection, preparation, analysis, and reporting; establishing more systematic links between the existing body of research in the social sciences; and moving away from proofs-of-concepts toward the systematic development and testing of hypotheses.

With this special issue, we want to contribute to the debate on how to use digital trace data in the social sciences productively. To us, this means not restricting one's research to online phenomena or succumbing to the temptation to use digital trace data to draw inferences on unrelated phenomena. We are very happy that we were able to collect five contributions from a multidisciplinary team of

CONTACT Andreas Jungherr andreas.jungherr@gmail.com Assistant Professor for Social Science Data Collection and Analysis, University of Konstanz, Universitätsstraße 10, D-78464 Konstanz, Germany.

scholars, demonstrating the potential of digital trace data in the social sciences, without becoming apologists for an uncritical empiricist approach to social science research.

Before we come to the contributions themselves, we will use this introduction to sketch the promises associated with the use of digital trace data in the social sciences. We will then continue with a discussion of the challenges emerging from the use of digital trace data. Before introducing the articles collected in this special issue, we will identify patterns evident in some of the most prominent work with digital trace data on topics related to politics. We will close with a short outlook on perspectives for further research and necessary development of the field.

The empiricist's promise

The growing use of digital services and devices in everyday life has created a data deluge. This deluge comes in the form of digital trace data—data wittingly or unwittingly produced in the context of using digital services or devices (Howison, Wiggins, & Crowston, 2011). These data hold great promise for business, government, and academia as they potentially document individuals' behavior unfiltered by obtrusive or unreliable measurements—such as self-reported behavior or observations in artificial laboratory environments. Additionally, these data come at a scale previously unknown to social scientists as they potentially document the behavior of each user of a given service or device. Finally, these data provide an impressive level of granularity in potentially documenting each user's every interaction with a digital service. Precision, size, and depth are thus three features of digital trace data widely perceived as carrying vast promise in the social sciences by constituting a "measurement revolution" (Watts, 2011). This promise has become widely associated with the terms "computational social science" and "big data" (Alvarez, 2016; Golder & Macy, 2014; Lazer et al., 2009; Lazer & Radford, 2017; Schroeder, 2016; Strohmaier & Wagner, 2014).

By offering a precise and unobtrusive documentation of user behavior, digital trace data potentially allow social scientists to avoid three traps of other measurement approaches (Golder & Macy, 2014; Howison et al., 2011; Salganik, 2017). For researchers interested in respondents' behavior and attitudes, reliance on self-reported behavior—as in surveys—is notoriously problematic. Too strong are potential measurement biases introduced by respondents' weak memory regarding behaviors of interest (Tourangeau, Rips, & Rasinski, 2000, pp. 82–92) and by their conscious or subconscious attempts at presenting socially acceptable answers (DeMaio, 1984). This is especially relevant for work in political communication and public opinion (Prior, 2009). Another benefit of this characteristic of digital trace data is that they emerge from user behavior in a natural setting. They thereby avoid potential biases emerging from artificial conditions researchers would have to put their subjects in to observe behavior (Levitt & List, 2007a, 2007b).

The size digital trace data tends to come in also matters (Golder & Macy, 2014; Salganik, 2017). The size of data sets allows researchers to compare patterns between subgroups in populations or categories, thereby making visible behavioral patterns of subgroups too small to be examined in survey samples and allowing meaningful comparisons between subcategories with small relative weight. Also, the size of digital trace data might allow the identification of small effect sizes otherwise indistinguishable from noise.

Finally, digital trace data offer a very fine-grained and detailed look at user behavior over time (Golder & Macy, 2014; Lazer et al., 2009), as well as interactions between different users and their broader consequences for political or communication processes. This allows researchers a detailed look at even minute behavior at the individual level, temporal developments over time, and the dynamics of previously invisible interactive exchanges. This promises social scientists a window into up-until-now invisible social phenomena and events.

Taken together, these characteristics make unlocking the potential of digital trace data in the social sciences of great importance:

> … just as the invention of the telescope revolutionized the study of the heavens, so too by rendering the unmeasurable measurable, the technological revolution in mobile, Web, and Internet communications has the potential to revolutionize our understanding of ourselves and how we interact. (Watts, 2011, p. 266)

In this, the debate about the scientific potential of big data and computational social science puts heavy

duty on the empirical potential of new data sources, be it by equating the potential of such traces for the social sciences with the invention of the "telescope" for astronomy (Watts, 2011), or by putting forward hopes for the identification of a "social physics" (Pentland, 2014). In these accounts, there is a strong empiricist undercurrent. Social science before big data appears as inherently soft and unreliable when compared to the natural sciences; a state of affairs that an increase in data available to researchers would remedy, thereby finally transforming the social sciences into a "true" science. In these accounts, increases in knowledge are only hampered by practical issues such as data access, data storage, user privacy, method development, or the dissemination of method training among graduate students (Golder & Macy, 2014; King, 2011; Lazer et al., 2009). Not surprisingly, this early enthusiasm has been met with critique.

The empiricist challenged

The optimism associated with the use of digital trace data tends to somewhat overshadow the very real challenges in making these new data sources actually work for social scientists. For all the diagnosed potential and promise of these new data sources, most research using digital trace data has as of yet fallen short of producing findings robustly contributing to central debates in the social sciences. This is not to take away from the inspired work based on digital trace data. But when examined closely, very few studies go beyond either illustrating usage patterns of digital services and devices or providing imaginative proof-of-concept type studies illustrating potential uses of digital trace data. In any case, the promised "measurement revolution" has not yet translated into a knowledge revolution for the social sciences. One reason for this is technical; using digital trace data in the social sciences simply is harder than early enthusiasts made us believe. The other is conceptual: there has to be a conscious effort in determining how to link digital trace data with sophisticated readings of contemporary theoretical debates in the social sciences. Whereas the first challenge is receiving increasing attention, the second is addressed only seldom.

In the discussion on how to use digital trace data for research, technical questions feature strongly. Front and center here is the question of what data one is actually given access to. One of the early tenets of big data research has been that these new data sources would make sampling superfluous as one would have access to every data point of every individual of relevance. This claim has been neatly labeled "n = all" (Mayer-Schönberger & Cukier, 2013, p. 26). Although popular, this early claim has turned out to be a fallacy (Jungherr, 2017). Digital trace data available to researchers are far from complete. Instead, it is best to think of any set of digital trace data as the result of various selection steps leading to the creation of a sample of unknown relationship to an original data set and the phenomenon of interest.

On a fundamental level, digital trace data provide researchers only with a slice of human behavior, namely, that which happens on and through the service or device that provided the data. Thus, these data can be used only as indicators with regard to aspects of political phenomena closely associated with the uses of digital services or devices providing the data; all aspects of social life going beyond this slice remain invisible. This limitation might be of little consequence if researchers focus on identifying usage patterns on selected platforms. But once researchers try to draw inferences on larger social phenomena than those directly related to their platform of focus this limitation becomes crippling (Jungherr, 2017; Salganik, 2017). Who can confidently claim that the patterns found in a set of digital trace data speak of more than simply specific usage practices associated with the affordances and usage culture of said service?

Yet, even with regard to the slice of human behavior visible in data traces of a specific service, researchers cannot in good conscience claim to have access to all data. Access to digital trace data depends on the provisions of corporations holding the data in the first place. As a rule, these are enterprises whose policies on data storage, retention, and access provision for third parties—such as researchers—follow commercial, operational, and legal considerations. This, again, leads researchers to gain access not to everything a user did on a given platform, but only to a selection of potentially relevant behaviors determined by the access policies, themselves determined by the corporation holding the data. In practice, this either happens through application programming interfaces (API), Web scraping, or by data dumps

for privileged partners. All three access methods are problematic because the relationship of the collected data and the underlying complete data set can only be guessed at (Jungherr, 2017; Salganik, 2017).

Another issue arises from the fact that the active user base of digital services is not representative of the general population (Blank, 2016; Hargittai, 2015). Again, this might matter little if we are only interested in examining usage patterns on specific digital services. But if we want to use digital trace data to draw inferences on the general population, this systematic skewness becomes very problematic. This is reinforced as the composition of the active user base of digital services, such as Facebook and Twitter, for example, seems to fluctuate unforeseeably over time and in reaction to specific events (Diaz, Gamon, Hofman, Kiciman, & Rothschild, 2016). This makes the development of weighing procedures that would account for a stable skewness between the users of digital services and the general population unfeasible. Accordingly, public opinion scholars have warned against using digital trace data uncritically as proxies for other data sources (Murphy et al., 2014; Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016). Add to these issues the very real challenges emerging from establishing a robust workflow with digital trace data in the social sciences (Freelon, 2015; King, 2011), and you find that the empiricist's promise seems much more elusive than originally thought.

Much less discussed but at least as important are conceptual issues in the work with digital trace data. In working with digital trace data, researchers usually do not explicitly discuss how their chosen data traces correspond with their phenomenon of interest. Instead, researchers simply assume their data to validly represent the phenomenon at the center of their work. This has been called the "mirror fallacy" (Jungherr, 2017). Digital data traces are in fact the result of a complicated mediation process. Phenomena of interest are filtered by individual-level factors—such as interests, preferences, and usage motives—and technological features of the service in questions—such as its code, algorithms, and affordances (Jungherr, 2015; Jungherr, Schoen, & Jürgens, 2016a). Ignoring these mediating factors in the interpretation of patterns identified in digital trace data means ultimately not being able to differentiate their impact from the imprint of social phenomena.

After accounting for the influence of mediating factors in the production of digital trace data, researchers also have to pay attention to linking the signals identified in the data to phenomena of interest. Establishing a valid link between data and concepts is of central importance in social science (Lazer, 2015). Establishing this link has, as of yet, not been at the center of work based on digital trace data (for notable exceptions see Casas & Williams, 2016; Fazekas, Popa, Schmitt, Barberá, & Theocharis, 2017; González-Bailón & Wang, 2016; Jungherr, 2017; Jungherr, Schoen, Posegga, & Jürgens, 2016b; Theocharis, Barberá, Fazekas, Popa, & Parnet, 2016). This relative neglect might be acceptable as a sign of an early stage in the development of the field when the emphasis of researchers is necessarily on first establishing the research potential associated with a new data source. But this neglect of at least attempting to seriously link signals identified in digital trace data to sophisticated readings of concepts in the social sciences leads to highly vulnerable research with dubious connections to debates in the social sciences.

Ultimately, these are questions of interpretation. Digital trace data do not speak for themselves. The empiricist's hope of simply looking at the data and identifying laws of social life is misplaced. Signals identified in digital trace data have to be linked to concepts and mediating factors have to be accounted for. Without this theoretical link, the empiricist's promises will remain unfulfilled (Gerbaudo, 2016; Jungherr, 2017).

Thus, the empiricist has been challenged in realizing the promise of the "measurement revolution" for the social sciences with regard to technical and methods-based aspects as well as conceptual and theoretical questions. But what can we learn from actual studies in political science using digital trace data. How has the empiricist's promise been realized?

The promise realized?

Whereas the work with digital trace data touches on various areas in the social sciences, our focus is on topics related to politics, an area for which digital trace data hold much promise and relevance (Alvarez, 2016; Clark & Golder, 2015; Freelon, 2015; Gil de Zúñiga & Diehl, 2017; Jungherr, 2015; Shah, Cappella, & Neuman, 2015). Here is not the place for a systematic review of research using digital

trace data in political science and related fields. Instead, we would like to offer a brief account of some central tendencies in work speaking to politics using digital trace data. From this account, it will become evident if and how the empiricist's promise has come to be realized in political science.

If we look at the available literature using digital trace data in the analysis of political phenomena, studies tend to fall into one of two categories. For one, we have an evolving segment of research using digital trace data to illustrate online components of political events, such as protests, televised debates, or election campaigns (Jungherr, 2014, 2015; Jungherr & Jürgens, 2014; Lotan et al., 2011; Nulty, Theocharis, Popa, Parnet, & Benoit, 2016; Wells et al., 2016). The trouble with studies largely focusing on simply describing political online phenomena without establishing larger context is that they can only be a first step in increasing our understanding of politics online. Although the best of these studies tell us a lot about how specific digital tools are used, they tell us very little about how theses uses matter for politics at large (some of our own work is very much vulnerable to this charge). Here, we often find that studies are highly imaginative in the use of data but spend inadequate time on linking signals meaningfully to concepts. This makes them very impressive exercises in advanced analytical procedures, but in linking the results of their studies to politics at large authors predominantly resort to storytelling.

Central to this development are conceptual questions. Which political phenomena of interest can be expected to be represented in digital trace data? In other words, which aspects of politics at large can be found in digital traces and which will remain invisible? How do politics online and offline intersect and what can we say about this process using digital trace data? After developing adequate concepts, identifying and testing mechanisms is a logical next step. Finally comes the question of measurement. How can we reliably measure political phenomena by using digital trace data? How do we develop procedures for indicator validation to make sure signals identified by us actually reliably and validly measure our phenomena of interest (Jungherr, 2015; Jungherr et al., 2016a, 2016b).

A helpful example comes from work focusing on digital politics. At the center of work on digital politics are questions of how to develop concepts adequately covering two important democratic processes: political participation and mobilization. What qualifies as online participation? Does it differ from offline participation? Should we be distinguishing between digitally enabled and digitally enhanced participation, and who pursues each of them? Does online participation fundamentally change mobilization and, consequently, collective action dynamics? Can digital media replace organizations in their fundamental role of organizing and mobilizing the public, and what is their democratic outcome? These are just some of the important questions scholars have been focusing on, giving rise to different theories of participation and collective action organization. Bimber, Flanagin, and Stohl (2012), Earl and Kimport (2011), Bennett and Segerberg (2013), and Shirky (2008), among others, have made important theoretical contributions about the changing nature of collective action in the era of digital media that are leading to the rethinking of organization and mobilization dynamics.

Although these concepts lend themselves to new types of investigations with digital trace data, they raise the important question of measurement. How can we develop standards to measuring relevant aspects of politics online? What has to be recorded and how can this be done reliably? Thinking about the changing nature of collective action has produced a number of rich, theory-driven contributions that utilize digital trace data, and that attempt to translate complex concepts into measurable phenomena. For example, Earl, Hurwitz, Mesinas, Tolan, and Arlotti (2013) have used Twitter data to examine how digital media alter traditional informational asymmetries between protesters and policemen (which could fundamentally change collective action dynamics), and Casas and Williams (2016) have used thousands of images sent on Twitter to assess whether images have a fundamental mobilizing role. González-Bailón and Wang (2016), Barberá et al. (2015), Theocharis, Lowe, van Deth, and García-Albacete (2015), and Sajuria, VanHeerde-Hudson, Hudson, Dasandi, and Theocharis (2015) have used a variety of text- and network-based approaches to operationalize connective action on Twitter and better understand the much-debated role of peripheral users and that of organizations in different types of collective action. Finally, Freelon, McIlwain, and Clark (2016) have developed metrics for measuring three theoretically

grounded metrics of social movement power: unity, numbers, and commitment.

Work on such questions comes with additional pitfalls. For example, while working on new phenomena of political participation and mobilization online, there is always the risk of overestimating their actual impact. Naturally, work tends to focus on interesting and popular cases, which is why, for example, there is so much work on the Occupy movements in all their manifestations across the world. But what makes these cases interesting might make them unrepresentative of the use of digital tools in politics as a whole. Another issue might arise from focusing on cases in which digital tools played a decisive role in politics at large. Although these cases might be informative, they are also likely to represent fringe phenomena. By perpetually focusing on the flavor of the month, researchers thus run the danger of losing sight of the laws governing the predominant share of political processes and phenomena. These issues are of course not specific to the work with digital trace data but are an inherent challenge for case selection in the social sciences (Gerring, 2012, 2017). But the perpetual exceptionalism that proponents of the work with digital trace data tend to frame their manifestos with, and the very strong influence of scientific protocols from computer science, engineering, and the natural sciences, tend to obscure these pitfalls. This makes it all the more important to restate the importance of social science methodology in addressing political phenomena. To exaggerate, instead of falling for an often highly simplistic approach to researching political or social phenomena driven by approaches based in computer science, engineering, or the natural sciences it seems more promising to "normalize" the work with digital trace data by anchoring it firmly in pluralistic methodological traditions in the social sciences (Jungherr, 2017). This means explicitly addressing the place of theory in the work with digital trace data.

In the other group, we find studies using digital trace data as predictors of political events and phenomena or proxies for other more traditional measurement approaches in the social sciences, such as surveys (Barberá, 2015; DiGrazia, McKelvey, Bollen, & Rojas, 2013; Steinert-Threlkeld, Mocanu, Vespignani, & Fowler, 2015; Tumasjan, Sprenger, Sandner, & Welpe, 2010). Here, we find a strong prominence of studies concentrating on statistical predictions of various political outcomes based on signals found in digital trace data (see Hofman, Sharma, & Watts, 2017; Schoen et al., 2013). In style and design, these studies follow computer science papers attempting to predict social phenomena or economic outcomes based on digital trace data (e.g., Choi & Varian, 2012). Although this literature is often very sophisticated in the way data are collected, analyzed, and modeled, at its core these contributions seem deeply uninterested in establishing the nature of the link between the variables in their model. On the one hand, this makes these studies highly popular in that they seemingly offer a fairly straightforward way to measure and predict social, economic, and political phenomena. On the other hand, these studies have been found to be highly vulnerable to replication efforts, indicating that early hopes might be a case of collective overexcitement rather than the hoped-for replacement for more traditional measurement approaches or the prediction of future developments (Gayo-Avello, 2013; Jungherr, Jürgens, & Schoen, 2012; Lazer, Kennedy, King, & Vespignani, 2014; Metaxas, Mustafaraj, & Gayo-Avello, 2011). Here, it is of paramount importance that social science loses its fascination with the proof-of-concept publication model imported from computer science and instead demands sophisticated tests of indicator validation and the theorizing and testing of links between variables ostensibly linked. Otherwise, social science will be drowned in false positives—ultimately ill-founded claims of digital trace data predicting social and political phenomena based on single-shot case studies (Jungherr, 2017; Jungherr et al., 2016b).

What is missing for the most part in the literature are studies connecting digital trace data meaningfully to central debates in the social sciences, or meaningfully extending our conceptual framework to account for political phenomena online. This lack means that, with some prominent exceptions, current research has little to say on the true social or political impact of digital tools, or on how to advance current social science debate. One cannot help but feel that a more conscious anchoring of work based on digital trace data in theory would help. Here, three areas are especially promising. First, we have to think more actively about establishing a measurement theory for digital trace data. How can we establish categories and procedures linking signals found in digital trace data validly to conceptualizations of political and social

phenomena of interest (Jungherr, 2017; Jungherr et al., 2016a)? For this it pays to disregard the exceptionalism surrounding digital trace data and instead connect with methodological work in statistics (e.g., Donoho, 2015; Efron & Hastie, 2016). The establishment and continuous work on such a measurement theory would do much in allowing us to more stably establish digital trace data in the social science mainstream.

Correspondingly, work based on digital trace data could also contribute to theory building. Although establishing links to existing theory is promising, the increasing digitalization of social and political life offers many chances to rethink the established body of theories (Neuman, 2016). Which concepts and mechanisms still have merit and which have overstayed their historical moment? How are digital tools changing communication and politics as we knew them? Here, digital trace data promise an interesting perspective. An instructive example is the bulk of work that has focused on challenging the well-established contours of the classic theory on the logic of collective action (Olson, 1965). These are questions of high social importance. Take, for example, the extensive debate about the perceived political power of misinformation online—so-called fake news. By using digital trace data to identify usage patterns and effects, and by providing theories linking these empirical patterns to social or political phenomena of interest, social scientists can further scientific debate on politics in contemporary societies as well as provide advice in the attempts to solve central social concerns (Watts, 2017). For example, in a recent, highly imaginative study on an equally widely debated current issue, trolling and online harassment, Munger (2016), using Twitter data, developed a method for reducing the use of anti-Black racist slurs by White men on the platform, advancing the study of prejudiced behavior and presenting a new way for battling online harassment.

All this shows that realizing the potential of digital trace data for the social sciences takes work on various issues. We need a focused debate on the practicalities of collecting and analyzing digital trace data. This is a debate that is indeed productively unfolding. Less prominent but of at least as much importance is the linking of signals found in digital trace data to concepts of interest. This search for a measurement theory of digital trace data in the social sciences is only beginning. Predominantly ignored is the final aspect, the meaningful connection of patterns identified in digital trace data with central debates in the social sciences or the establishment of new theories accounting for communication and politics online. This constitutes the empiricist's challenge. How to connect empirical evidence meaningfully so that a larger picture of the nature, patterns, and mechanisms of contemporary political life on, and with, digital tools can emerge. We are happy to be able to present five strong responses to this challenge.

The articles in this special issue

This special issue collects five contributions, all of which use or address social media data, one of the most prominent sources of digital trace data in the social sciences. In many ways, these contributions demonstrate the best of research with digital trace data and allow us to extract four *motifs*, highlighting a selection of promising features and practices.

The first motif is *contextualization and theoretical framing*. One of the most consistent critiques of research based on digital trace data is the relative neglect of contextual framing and theoretical linking in favor of an overwhelming emphasis on the data themselves, their quality, and their magnitude (boyd & Crawford, 2012; van Dijck, 2014). Without exception, the authors in this special issue demonstrate how theory-driven approaches can lead to innovative investigations into the empirical applicability of prominent theoretical concepts while, at the same time, providing conceptual tools and measures others can build on. To name just two examples, in his article, Freelon offers an operationalization of Stromer-Galley's "controlled interactivity" concept of electoral campaigning (Stromer-Galley, 2014), thereby proposing specific empirical criteria that can be used to test its efficacy. Spaiser and her colleagues draw on Castell's theory of "communication power" (Castells, 2009) to study the online struggle for power during the 2011–2012 Duma and presidential elections between pro-government and oppositional groups in Russia.

The second motif is the relationship between *data and media affordances*. Although, as Karpf (2017, p. 5) notes, we still tend to think of "the media system" as mainly characterized by the broadcast media institutions that dominated the twentieth century, in the

hybrid media system (Chadwick, 2013) the structure, routines, and operations of these same institutions have been reconfigured (if not replaced) in reaction to pressures by new media that play an increasingly powerful role in setting the agenda (Jungherr, 2014). In their search for power, individual and institutional actors (such as candidates, social movements, or political parties), adjust their strategies on the basis of what different platforms afford them. These new parameters imply that the study of politics—especially political communication—can hardly rely on singular data sources any longer because this would imply missing important pieces of the puzzle, given that different players may find that a certain (or a combination of) media environment affords their strategic purposes better than others. Thankfully, the rise of the hybrid media system coincides with the rise in the availability of digital trace data as well as the computational capacity to analyze them. In the articles included in this special issue, we observe both highly fruitful syntheses of different data sources and clear understandings regarding why sampling these sources according to the interests of each study is more helpful for researchers than attempting to get "everything that's out there." In their study of how Twitter is used as a tool for political communication, McGregor and colleagues combine three million tweets from three distinct groups (news media, political actors, and the public—a choice reflecting the changing balance of power in the political communication environment), with publicly available data on candidates and elections as published in the news media, the Federal Election Commission, and other reports. Aiming to capture the Russian government's efforts to defame, discredit, marginalize, and countermobilize the opposition in an environment in which the opposition can (presumably) react more swiftly and visibly than in highly controlled traditional media such as TV or radio, Spaiser and colleagues analyze more than 700,000 public tweets, allowing them to carry out a dynamic discourse analysis. Facebook's huge user base and highly configurable interactive environment that affords candidates great control over what is posted on their Facebook pages, provides Freelon with the ideal platform to operationalize and empirically measure controlled interactivity. Similarly, in their study of candidate-to-candidate interaction, Laaksonen and colleagues use a data set of 137,000 Facebook contributions from 1,111 Finish candidate pages with the aim of studying interactions in a media environment that can accommodate in-depth conversations.

As a third motif, we distinguish what can be probably best referred to as *innovative methodological plurality*. As King has argued, "big data" should not be predominantly about the data. Instead, our focus should be on developing innovative analytical methods in response to research opportunities provided by these data (King, 2016). This special issue demonstrates the benefits of this approach, not only through the use of innovative analytical methods, but through the manifestation of methodological plurality. Digital trace data may have posed considerable challenges for social science research, but as social scientists from different methodological standpoints have come together to face this challenge, new ideas about how combinations of qualitative and quantitative approaches can be united with the goal of shedding light on both old and new questions are emerging. In one such attempt that combines participatory observation and big-data analysis, Laaksonen and colleagues put forward their manifesto on how what they call "big-data-augmented ethnography" can enhance our understanding of candidate-to-candidate interactions, outlining at the same time the different conceptual and empirical stages of their proposed approach. By flicking through the pages of this special issue, the reader can also learn more—and assess the benefits of—innovative methods in the work with digital trace data. Starting from traditional statistical applications that assess the relationship between the volume of Twitter data and vote share (see the article by McGregor and colleagues), continuing with follower-based network analysis (see the article by Spaiser and colleagues), and several text analytic approaches—such as dictionary-based methods for examining candidate interactions on Facebook (see the article by Freelon), and word count, n-gram, and sentiment analysis for identifying pro-Putin and opposition Twitter users in an attempt to understand discourse dynamics in opposing electoral camps (again, Spaiser and colleagues)—the articles present a variety of methods in service of addressing specific research questions.

The fourth and last motif, *interdisciplinary approaches*, refers not to the work published in this special issue but to the collaborative effort on the part of the scholars behind it. One of the objectives in

proposing the special issue was to bring together an interdisciplinary group of scientists addressing in their different fields the challenges of working with digital trace data. It was our impression that although many of the challenges were common across disciplines, an interdisciplinary conversation about how to overcome them was lacking. As mentioned, the challenges of these new data sources are technical, conceptual, theoretical, and philosophical and although some disciplines might be well equipped to deal with some of these challenges, each single one is very limited in addressing others. Although the popularization of digital trace data and the related challenges have clearly narrowed this divide, it is with great pleasure that we see that the result of the work published in this special issue is, collectively, the outcome of political scientists, media and communication scientists, computational social scientists, computer scientists, and speech communication specialists.

The four motifs summarize a number of characteristics found in the papers published in this special issue. We believe they can support the scholarly community, not only in its continuous effort to ask and answer meaningful questions in the age of big data. Here, it is important to not lose sight of the broader, societal consequences of the availability of massive amounts of data. In an article discussing the consequences of data-based surveillance, one of the major questions posed by the emergence of big data, Nick Couldry reflects on both the possibilities and perils of this phenomenon for democracy more broadly. The rise of what Rainie and Wellman (2012) have called a "networked social operating system" over the last five decades has led to shifts in the infrastructure of communication, and has given rise to the emergence of entirely new ways for citizens to connect with one another through their technologies of choice. Shifts such as these, Couldry argues, are changing fundamentally the nature of institutional power. In his article "Surveillance Democracy" he discusses how the rise of an infrastructure of surveillance brings new tensions to our notions of autonomy and freedom in democratic societies. Pointing out that data-driven processes with continuous tracking and categorizing functions have been embedded in spaces of individual autonomy and social interaction today, Couldry asks if we have reflected sufficiently on the potential costs of a potential remodeling of democracy and social and political life.

Asking meaningful questions in the age of big data

As is often the case with technological phenomena forcefully entering the public debate, the conversation surrounding the potentials and risks of digital trace data has been characterized by much hyperbole. This can be clearly seen in the use and the debate surrounding the term "big data." Some have excitedly praised their capacity to revolutionize every aspect of contemporary life, while others have warned about the hidden dangers and risks. These extreme positions have somewhat more moderately found expression in the discussion on the uses of digital trace data in political science. Here, their potential was welcomed by some as an opportunity to overcome limitations of other more traditional measurement approaches, while for others they posed a threat encouraging inductive, predominantly data-driven research divorced not just from established theories, but also from its ostensive topic, politics. In practice, neither the highly optimistic, nor the darkly pessimistic expectations have been realized. Yet, elements of both are evident in contemporary scholarship, raising questions about how to realize the potential of this new data source in addressing meaningful questions while remaining conscious of their challenges.

Even when acknowledging the many pitfalls in the work with "big data," we believe their potential for the social sciences in general and political science in particular is considerable. As many have correctly noted, digital trace data offer us the opportunity to better understand social and political phenomena viewed through a new lens. That this lens comes with specific distortions and blind spots does not mean we should discard it. Instead, we should spend time and effort to understand these specific flaws so that we can account for them in the interpretation of data made available to us through the use of this new data type. This means putting a stop to exaggerated expectations, seeing in digital trace data a one-stop solution for everything that ails social science or as the final key to unlocking natural laws of social life. Instead, we should approach this data source as any other, with a measurement

theory informing us on the interpretation of findings based on digital trace data and through research designs accounting for the specific potentials and limitations of digital trace data (Jungherr, 2017; Salganik, 2017).

Most of all, we believe we should focus on how digital trace data allow us to ask new and meaningful questions in the social sciences instead of losing ourselves in ultimately fruitless games of prediction. This implies: (a) using these new data to reassess existing theories, but most importantly building new ones in light of new insights that could not have been acquired with previous research tools; (b) developing new concepts and measures that, in combination, can help us better understand how attitudes and behaviors captured by this new data source map not only onto larger phenomena, but also onto our existing understandings, thereby making clearer what inferences we can—and cannot—draw in the study of complex social and political processes; (c) reassessing our epistemological tools and methods and through interdisciplinary collaborations, reattuning them to synergize, rather than compete, with one another; and (d) making sure that this entire research program remains consistent with scientific values, ethics, and practices.

Research using digital trace data has made huge strides in a very short time. As discussed, this new data source enables remarkable and exceptionally creative scholarly work, asking meaningful questions and providing insightful answers about the workings of a variety of social and political phenomena and processes, all the while inviting us to revisit established theories and revitalizing interest in social and political processes whose mechanics social scientists presumed as well established for more than half a century. In light of this ongoing development, we are excited to present you with a stimulating collection of articles that demonstrate some of the best aspects of big-data research and at the same time provide thrilling new insights into various topics in the field of information technology and politics.

Acknowledgments

This special issue is an offshoot of a conference we organized at the Mannheim Centre for European Social Research (MZES) in autumn 2015. We thank the MZES and the Lorenz-von-Stein-Gesellschaft for graciously providing funding for this event. Also, we want to thank the keynote speakers—W. Lance Bennett, Sandra González-Bailón, Jonathan Nagler, and Richard Rogers—and the participants for providing their valuable perspectives and getting us started on a road that led us to the arguments presented in these pages.

Notes on contributors

Andreas Jungherr is an assistant professor of Social Science Data Collection and Analysis at the University of Konstanz, Germany. His research focuses on the impact of digital technology on political communication and the use of digital trace data in the social sciences. His research has been published in *Journal of Communication, Journal of Computer-Mediated Communication, The International Journal of Press/Politics*, and *Social Science Computer Review*.

Yannis Theocharis is a research fellow at the Mannheim Centre for European Social Research (MZES). His research interests are in online and offline political participation, digital media, social capital, and social network analysis. His research has been published in the *Journal of Communication, New Media & Society, Electoral Studies*, and *European Political Science Review*.

ORCID

Andreas Jungherr http://orcid.org/0000-0003-2598-2453

References

Alvarez, R. M. (Ed.). (2016). *Computational social science: Discovery and prediction*. New York, NY: Cambridge University Press.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91. doi:10.1093/pan/mpu011

Barberá, P., Wang, N., Bonneau, R., Jost, J. T., Nagler, J., Tucker, J., & González-Bailón, S. (2015). The critical periphery in the growth of social protests. *Plos One*, *10*(11), e0143611. doi:10.1371/journal.pone.0143611

Bennett, W. L., & Segerberg, A. (2013). *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge, England: Cambridge University Press.

Bimber, B., Flanagin, A. J., & Stohl, C. (2012). *Collective action in organizations: Interaction and engagement in an era of technological change*. Cambridge, England: Cambridge University Press.

Blank, G. (2016). The digital divide among Twitter users and its implication for social research. *Social Science Computer Review*. doi:10.1177/0894439316671698

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Casas, A., & Williams, N. W. (2016). *Images that matter: Online protests and the mobilizing role of pictures*. Paper

presented at the 112th Annual Meeting of the American Political Science Association, Philadelphia, PA.

Castells, M. (2009). *Communication power*. New York, NY: Oxford University Press.

Chadwick, A. (2013). *The hybrid media system: Politics and power*. New York, NY: Oxford University Press.

Choi, H., & Varian, H. A. L. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2–9. doi:10.1111/j.1475-4932.2012.00809.x

Clark, W. R., & Golder, M. (2015). Big data, causal inference, and formal theory: Contradictory trends in political science? *PS: Political Science & Politics*, *48*(1), 65–70. doi:10.1017/S1049096514001759

DeMaio, T. (1984). Social desirability and survey measurement: A review. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 257–281). New York, NY: Russell Sage Foundation.

Diaz, F., Gamon, M., Hofman, J. M., Kiciman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *Plos One*, *11*(1), e0145406. doi:10.1371/journal.pone.0145406

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *Plos One*, *8*(11), e79449. doi:10.1371/journal.pone.0079449

Donoho, D. (2015). *50 years of data science*. Paper presented at the John W. Tukey 100th Birthday Celebration at Princeton University, Princeton, NJ. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Earl, J., Hurwitz, H. M., Mesinas, A. M., Tolan, M., & Arlotti, A. (2013). This protest will be tweeted: Twitter and protest policing during the Pittsburgh G20. *Information Communication & Society*, *16*(4), 459–478. doi:10.1080/1369118x.2013.777756

Earl, J., & Kimport, K. (2011). *Digitally enabled social change: Activism in the Internet age*. Cambridge, MA: MIT Press.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge, England: Cambridge University Press.

Fazekas, Z., Popa, S. A., Schmitt, H., Barberá, P., Theocharis, Y., & Parnet, O. (2017). *Issue politicization on social media: Addressing emerging issues in election campaigns*. Working Paper. Retrieved from https://zfazekas.github.io/papers/eu_polit.pdf.

Freelon, D. (2015). On the cutting edge of big data: Digital politics research in the social computing literature. In S. Coleman & D. Freelon (Eds.), *Handbook of digital politics* (pp. 451–472). Northampton, MA: Edward Elgar Publishing.

Freelon, D., McIlwain, C., & Clark, M. (2016). Quantifying the power and consequences of social media protest. *New Media & Society*. doi:10.1177/1461444816676646

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, *31*(6), 649–679. doi:10.1177/0894439313493979

Gerbaudo, P. (2016). From data analytics to data hermeneutics. Online political discussions, digital methods and the continuing relevance of interpretive approaches. *Digital Culture & Society*, *2*(2), 95–112. doi:10.14361/dcs-2016-0207

Gerring, J. (2012). *Social science methodology: A unified framework* (2nd ed.). Cambridge, England: Cambridge University Press.

Gerring, J. (2017). *Case study research: Principles and practices* (2nd ed.). Cambridge, England: Cambridge University Press.

Gil de Zúñiga, H., & Diehl, T. (2017). Citizenship, social media, and big data: Current and future research in the social sciences. *Social Science Computer Review*, *35*(1), 3–9. doi:10.1177/0894439315619589

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, *40*(1), 129–152. doi:10.1146/annurev-soc-071913-043145

González-Bailón, S., & Wang, N. (2016). Networked discontent: The anatomy of protest campaigns in social media. *Social Networks*, *44*, 95–104. doi:10.1016/j.socnet.2015.07.003

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 63–76. doi:10.1177/0002716215570866

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. doi:10.1126/science.aal3856

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 767–797.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of Communication*, *64*(2), 239–259. doi:10.1111/jcom.12087

Jungherr, A. (2015). *Analyzing political communication with digital trace data: The role of Twitter messages in social science research*. Cham, Switzerland: Springer.

Jungherr, A. (2017). Normalizing digital trace data. In N. J. Stroud & S. McGregor (Eds.), *Digital discussions: How big data informs political communication*. Oxon, England: Routledge.

Jungherr, A., & Jürgens, P. (2014). Through a glass, darkly: Tactical support and symbolic association in Twitter messages commenting on Stuttgart 21. *Social Science Computer Review*, *32*(1), 74–89. doi:10.1177/0894439313500022

Jungherr, A., Jürgens, P., & Schoen, H. (2012). Why the pirate party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment." *Social Science Computer Review*, *30*(2), 229–234. doi:10.1177/0894439311404119

Jungherr, A., Schoen, H., & Jürgens, P. (2016a). The mediation of politics through Twitter: An analysis of messages posted during the campaign for the German federal election 2013. *Journal of Computer-Mediated Communication*, *21*(1), 50–68. doi:10.1111/jcc4.12143

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2016b). Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*. doi:10.1177/0894439316631043

Karpf, D. (2017). *Analytical activism: Digital listening and the new political strategy*. New York, NY: Oxford University Press.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–721. doi:10.1126/science.1197872

King, G. (2016). Preface: Big data is not about the data! In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (pp. vii–x). New York, NY: Cambridge University Press.

Lazer, D. (2015). Issues of construct validity and reliability in massive, passive data collections. *The City Papers: An Essay Collection from The Decent City Initiative*. Retrieved from http://citiespapers.ssrc.org/issues-of-construct-validity-and-reliability-in-massive-passive-data-collections/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. doi:10.1126/science.1248506

Lazer, D., Pentland, A., Adamic, L. A., Aral, S., Barabasi, A. L., Brewer, D., … Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Lazer, D., & Radford, J. (2017). Introduction to big data. *Annual Review of Sociology, 43*. doi:10.1146/annurev-soc-060116-053457

Levitt, S. D., & List, J. A. (2007a). Viewpoint: On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue Canadienne D'économique*, *40*(2), 347–370. doi:10.1111/j.1365-2966.2007.00412.x

Levitt, S. D., & List, J. A. (2007b). What do laboratory experiments measuring social preferences reveal about the real world? *The Journal of Economic Perspectives*, *21*(2), 153–174. doi:10.2307/30033722

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & boyd, d. (2011). The Arab Spring|The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, *5*, 1375–1405.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt.

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). *How (not) to predict elections*. Paper presented at the IEEE Third International Conference on Social Computing (SocialCom), Boston, MA. doi:10.1109/PASSAT/SocialCom.2011.98

Munger, K. (2016). Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*. doi:10.1007/s11109-016-9373-5

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., … Harwood, P. (2014). Social media in public opinion research: Executive summary of the aapor task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*(4), 788–794. doi:10.1093/poq/nfu053

Neuman, W. R. (2016). *The digital difference: Media technology and the theory of communication effects*. Cambridge, MA: Harvard University Press.

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social media and political communication in the 2014 elections to the European Parliament. *Electoral Studies*, *44*, 429–444. doi:10.1016/j.electstud.2016.04.014

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Pentland, A. (2014). *Social physics: How social networks can make us smarter*. New York, NY: Penguin Books.

Prior, M. (2009). Improving media effects research through better measurement of news exposure. *The Journal of Politics*, *71*(3), 893–908. doi:10.1017/S0022381609090781

Rainie, L., & Wellman, B. (2012). *Networked: The new social operating system*. Cambridge, MA: The MIT Press.

Sajuria, J., VanHeerde-Hudson, J., Hudson, D., Dasandi, N., & Theocharis, Y. (2015). Tweeting alone? An analysis of bridging and bonding social capital in online networks. *American Politics Research*, *43*(4), 708–738. doi:10.1177/1532673X14557942

Salganik, M. J. (2017). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly*, *80*(1), 180–211. doi:10.1093/poq/nfv048

Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*, *23*(5), 528–543. doi:10.1108/IntR-06-2013-0115

Schroeder, R. (2016). *Big data and communication research*. Retrieved from http://communication.oxfordre.com/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-276.

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. doi:10.1177/0002716215572084

Shirky, C. (2008). *Here comes everybody: The power of organizing without organizations*. New York, NY: The Penguin Press.

Steinert-Threlkeld, Z. C., Mocanu, D., Vespignani, A., & Fowler, J. (2015). Online social networks and offline protest. *EPJ Data Science*, *4*(19), 1–9. doi:10.1140/epjds/s13688-015-0056-y

Strohmaier, M., & Wagner, C. (2014). Computational social science for the World Wide Web. *IEEE Intelligent Systems*, *29*(5), 84–88. doi:10.1109/mis.2014.80

Stromer-Galley, J. (2014). *Presidential campaigning in the Internet age*. New York, NY: Oxford University Press.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The

consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–1031. doi:10.1111/jcom.12259

Theocharis, Y., Lowe, W., van Deth, J. W., & García-Albacete, G. (2015). Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*, *18*(22), 202–220. doi:10.1080/1369118X.2014.948035

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In M. Hearst, W. Cohen, & S. Gosling (Eds.), *ICWSM 2010: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media* (pp. 178–185). Menlo Park, CA: Association for the Advancement of Artificial Intelligence (AAAI).

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance and Society*, *12*(2), 197–208.

Watts, D. J. (2011). *Everything is obvious: How common sense fails us*. New York, NY: Random House.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behavior*, *1*. doi:10.1038/s41562-016-0015

Wells, C., Shah, D. V., Pevehouse, J. C., Yang, J. H., Pelled, A., Boehm, F., … Schmidt, J. L. (2016). How Trump drove coverage of the nomination: Hybrid media. *Political Communication*, *33*(4), 669–676. doi:10.1080/10584609.2016.1224416