Erschienen in: Public Administration Review ; 76 (2016), 6. - S. 928-937
https://dx.doi.org/10.1111/puar.12625

Ines Mergel
University of Konstanz, Germany

R. Karl Rethemeyer
University at Albany
Kimberley Isett
Georgia Tech

Big Data in Public Affairs

Abstract: *This article offers an overview of the conceptual, substantive, and practical issues surrounding "big data" to provide one perspective on how the field of public affairs can successfully cope with the big data revolution. Big data in public affairs refers to a combination of administrative data collected through traditional means and large-scale data sets created by sensors, computer networks, or individuals as they use the Internet. In public affairs, new opportunities for real-time insights into behavioral patterns are emerging but are bound by safeguards limiting government reach through the restriction of the collection and analysis of these data. To address both the opportunities and challenges of this emerging phenomenon, the authors first review the evolving canon of big data articles across related fields. Second, they derive a working definition of big data in public affairs. Third, they review the methodological and analytic challenges of using big data in public affairs scholarship and practice. The article concludes with implications for public affairs.*

Practitioner Points

- While "big data" refers to the scale of newly emerging data sets (many observations with many variables), the term also refers to the nature of the data collection process (continuous and automatic), the form of the data collected (structured and unstructured), the sources of such data (public and private), the "granularity" of the data (more variables describing more discrete characteristics of persons, places, events, interactions, and so forth), and the lag between collection and readiness for analysis (ever shorter).
- Big data in the public sector is context specific and needs to be meaningfully combined with administratively collected data to have value in improving public programs.
- There are important ethical issues, privacy concerns, security and secrecy problems, and feasibility and efficacy issues when using big data for the public good.

Public administration researchers and practitioners for most of the field's history have bemoaned the lack of data for analysis and operations. In the space of roughly two decades, the Internet has turned this problem on its head. Now, scholars and practitioners are scrambling to realize the opportunities and face the challenges that "big data" presents. These "big" data sets are increasingly used to help public managers derive real-time insights into behavioral changes, public opinion, or daily life. Additionally, researchers are using these data sets to validate existing theory and to generate new insights in areas in which few data resources previously existed and for which analytics are still under development (Chen, Chiang, and Storey Chen, Roger and Storey, 2012).

Despite the rhetoric surrounding these data, simply having access to and using algorithms for analysis of large-scale data sets does not necessarily lead to insights (Meier and O'Toole 2005). For instance, computer scientists often reveal the composition of large online networks but do not connect the findings to existing policy or public management frameworks (Eagle, Pentland, and Lazer 2009; Onnela et al. 2007). Indeed, much of the "promise" of these data has been their "post-theoretical" nature—focusing on the possibilities for discovery within huge and newly accessible data sets without well-developed conceptual foundations that also provide actionable insights for policy makers or public managers.

We seek to orient the field of public affairs to issues inherent in big data—especially those that are unique to public sector endeavors—and to the possible implications for theory and practice. Beyond the analytical and interpretative challenges, we see major hurdles for data collection, retention, and analysis of these types of data in the public sector. Current law and practices regulating individual-level data collection focus only on administratively collected data and do not easily extend to this new source of information.

We offer an overview of the conceptual, substantive, and practical issues surrounding big data in order to provide one perspective on how the field of public affairs can remain relevant as the innovative approaches of big data become ubiquitous. We have organized this article into four substantive sections. We first review the existing definitions and perspectives in neighboring fields, including public and social policy, management, and political science, as well as newer areas of study such as computational social sciences and policy informatics. The second section provides a definition of big data in public affairs and a critical discussion of what is currently missing in the literature. The third section examines a set of important issues with using big data that public affairs scholars must consider. We end this article with some thoughts on the implications of these new challenges and opportunities and offer an interpretive lens that allows us to highlight the comparative advantage that our field has in this endeavor.

Big Data Definitions and Perspectives in the Disciplines

"Big data" is currently used as an umbrella term to describe various aspects of this data-intensive approach. While the term "big data" is commonly used, the more precise terms of "data analytics" and "data science" have also been introduced as better descriptors of the phenomena—all are used interchangeably in this article. These terms simultaneously refer to the amount of data, computational practices used to harvest large-scale data sets from multiple sources, and analytical strategies that manipulate these data in real time. Data analytics focuses mostly on new forms of (social) data generated by Internet users, such as networks created through follower relationships on social networking sites, links between websites, or mobile phone connections and use of mobile apps that can be combined with the users' sociodemographic data. These data can also be generated by the "Internet of things"—devices that use the Internet to help control smaller, discrete activities such as the temperature of your house, the charge level in your electric vehicle, or other new technologies such as sensors (Bryant, Katz, and Lazowska 2008). Increasingly, devices also passively capture information about their owners and users such as location, schedule, speed, or health data.

Internet users, such as networks created through follower relationships on social networking sites, links between websites, or mobile phone connections and use of mobile apps that can be combined with the users' sociodemographic data.

Within public affairs, there are few published definitions of the concept that grapple with all of these dimensions. One exception is a White House report that uses a definition proposed by the

Table 1 Big Data Definitions across Disciplines

Discipline	Author(s)	Definitions	Opportunities	Challenges
Management	George, Hass, and Pentland (2014)	"Big data is generated from an increasing plurality of sources, including Internet clicks, mobile transactions, user-generated content, and social media as well as purposefully generated content through sensor networks and business transactions such as sales queries and purchase transactions" (321)	• Signaling functions to understand emerging vulnerabilities • Predict outcomes with greater precision	• Face-to-face communication versus automated analysis of behavioral patterns • Stated versus automatically detected preferences
Public policy	Pirog (2014)	New formats, quality, and availability of administrative data (volume, velocity)	• Completeness and changes in the types of data (Data.gov) • Real-time availability of data • Connecting biology, psychology, and public policy to study risky behavior • Geospatial data increasingly accessible through incorporation of geocodes in large social surveys	• Unstructured nature of the data • No breakthroughs in quasi-experimental research designs
Political science	Clark and Golder (2014)	"Technological innovations such as machine learning have allowed researchers to gather either new types of data, such as social media data, or vast quantities of traditional data with less expense" (65)	• Benefits for description and measurement • Access to "unfiltered" opinions	• Big data≠better research designs or≠causal inference
Information and technology management	Janssen and Van den Hoven (2015)	• BOLD—Big and Open Linked Data	• Create public value by combining and analyzing large-scale data sets	• Ethical, cultural, technological challenges
	Boyd and Crawford (2012)	• "massive quantities of information produced by and about people, things, and their interactions" (Janssen and Van den Hoven 2015, 662)		• Unresolved privacy intrusions
Computational social sciences	Lazer et al. (2009) Lazer et al. (2014) Denning (1990) Bryant, Katz, and Lazowska (2008)	"Second-by-second picture of interactions over extended periods of time, providing information about both the structure and content of relationships" (Lazer et al. 2009, 2)	• From individual-level data to society as a whole (micro to macro insights)	• Acquisition and storage of data • Design and test of algorithms • Detecting patterns • Overpredication/estimation of online searches (Lazer at al. 2014) • False interpretation of signals

National Science Foundation: "Big data sets are large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, e-mail, video, click streams, and/or all other digital sources available today and in the future" (White House 2014, 3). Scholars in neighboring fields have begun to operationalize the concept, offer definitions, and describe the use of data analytics more actively than those in public affairs (for an overview, see table 1). While there are volumes of articles on big data, we focus here on selected contributions in fields most closely and synergistically related to our own to provide scaffolding for our thinking.

As George, Haas, and Pentland (2014) highlight in the *Academy of Management Journal,* there is a misunderstanding in the way the term "big data" is often used. Especially in mainstream media and among practitioners, the term is used to describe the size of data sets, focusing mostly on hugeness rather than content or use (McAfee and Brynjolfsson 2012). Instead, they posit that "big data" is (ironically) about "granularity." That is, big data refers to the specificity of the data that can be collected and then combined with other sources to provide deep information about events, individuals, processes, or phenomena.

In political science, Clark and Golder refer to big data as "technological innovations such as machine learning [that] have allowed researchers to gather either new types of data, such as social media data, or vast quantities of traditional data with less expense" (2015, 65). They go on to state that "our increasing ability to produce, collect, store, and analyze vast amounts of data is going to transform our understanding of the political world." These scholars see the "big data revolution" as a challenge to their field to increase the sophistication and capacity of its data collection and analysis techniques. However, they do recognize that while "more data is better than less," access to large-scale data sets from the Internet and other sources is challenging the way political scientists think about research methods, data collection, and theory development. Big data also challenges scholars to think systematically about which data are relevant and which are chaff. Finally, Clark and Golder highlight how problematic the assumption is that increasing amounts of data automatically lead to better theory development or even improved research designs.

With respect to public policy, Pirog (2014) predicted in an editorial in the *Journal of Policy Analysis and Management* that the availability of new and more high-quality data sets, such as the Open Data offerings found on Data.gov, will transform experimental research in public policy. However, Pirog contradicts Clark and Golder, stating that these new data sets will *not* require changes in research design or statistical approaches. She instead asserts that big data only provides a more complete picture of individuals. In her view, understanding individuals better has the potential to improve public policy and public management research through improved specification of preferences and values. For Pirog, the key innovation is the availability of additional geospatial data, such as real-time satellite imagery and Global Positioning System (GPS) location data for cell phones, economic transactions, or Internet search data, that will—and here she quotes Paul Decker—create a "data tsunami" (Pirog 2014, 540). She views the challenges of big data as mostly attributable to its unstructured nature, thus needing to be combined through parsing methods that might be unknown to social scientists. As we will discuss later, an underlying challenge in using these data is the need to carefully select variables to avoid the "old" problem of overspecification.

In the information management and technology field, it is apparent that the big data discussion is closely related to the use of technologies that help link diverse data sources with each other. For example, Janssen and Van den Hoven (2015) focus on connecting administratively collected open data with Internet-generated information that does not follow formal design and collection processes. In the same vein, Boyd and Crawford (2012) focus on the opportunities that large-scale data collection and concatenation can provide to study social behavior, but they also predict that many ethical dilemmas, such as threats to privacy, are still unsolved and need to be addressed before public value is created.

Computer science considers, as Bryant, Katz, and Lazowska state, "big data computing as the biggest innovation in computing in the last decade" (2008, 7). These scholars are working on methods to retrieve and store data along with the design and testing of algorithms to detect structure and patterns in the data (Denning 1990). A major challenge within computer science is to uncover unexpected patterns and interpret, as well as act on, the insights from these discoveries (Frankel and Reid 2008). Another challenge is the lack of sufficient investment in data centers, high-capacity networking infrastructures, high-performance computing sites, such as national laboratories, supercomputers, or cluster systems (Doctorow 2008; Lynch 2008). While these challenges are important, they are of a different nature than those faced in public affairs with respect to data science.

These perspectives from across related fields highlight the need for cross-disciplinary collaboration among social scientists, who have substantive depth on research methods and theory, and computer scientists, who have the computational and methodological skills to construct and analyze algorithms on data structures discussed in this article. Two relatively recent fields have emerged that aim to address this need for new forms of interdisciplinary collaboration: computational social sciences (Lazer et al. 2009) and policy informatics (Johnston 2015). Lazer et al. highlight the ubiquitous nature of automatically generated social networking data through GPS tracking, video recording, or radio-frequency identification transmissions in public transit that result in insights about societies as a whole based on the combination of vast amounts of individual-level data. The key is "second-by-second picture(s) of interactions over extended periods of time, providing information about both the structure and content of relationships" (Lazer et al. 2009, 2). Johnston defines policy informatics as the "study of how computation and communication technology is leveraged to understand and address complex public policy and administration problems and realize innovations in governance processes and institutions" (2015, 1). He suggests that "computational methods"—which we consider to be both use of computerized technology to capture large-scale data sets in the first place and computation-intensive efforts to analyze those

data sets—are necessary to deal with the increasing governance complexities the public sector is facing.

While traditional disciplines and newer fields highlighted here have and should influence public affairs' engagement with big data, none fully considers the range of issues associated specifically with practice and research in public affairs.

Big Data Definition for Public Affairs

Big data in public affairs research focuses on the collection of multimodal digital data generated by public and private providers. While some aspects of big data are consistent across fields, at least one aspect is specific to ours. Starting with the similarities, big data includes (1) data created by private citizens through their interactions with each other online (such as social media data), and (2) data automatically generated from sensors in, for instance, buildings, cars, and streets, that is automatically transmitted online. However, public affairs also includes (3) data that are automatically collected by public entities in the course of their operations. Across mechanisms, both structured and unstructured data are collected (including metadata that describe attributes of those subjects and objects). The granularity of the data still remains relatively coarse today but is increasing as "Internet-enabled" devices (smart phones, Wi-Fi thermostats, car automation systems, etc.) proliferate and data systems are increasingly automated.

Big data in public affairs research focuses on the collection of multimodal digital data generated by public and private providers.

Big data is a moving target: what is possible now is less than what will be possible in the future. As we will discuss later, *today* big data provides insights about the mean distribution of general online preferences, energy consumption, movements, and so forth. For reasons we will detail, big data *today* does not necessarily allow analysts to derive insights about individual preferences or behaviors that are outside the mainstream.

However, one of the key promises of big data *tomorrow* is the ability—through sheer size and comprehensiveness—to analyze small populations, extreme outcomes, or rare events, that is, the "tails of the distribution." As device use, Internet participation, and the ability to build social science tools for data collection on the Internet grow, we should be able to study small, hard-to-discover subpopulations over time. The key, we believe, is carefully discerning the difference between what we can do *today* versus what we can do in the not-too-distant future. We also need to realize that there are issues with using large-scale, Internet-derived data sets that are irreducible and must be constantly considered.

One advantage that scholars using data analytic approaches have is the ability to examine problems in real time—or at least more nearly in real time as capabilities grow. Large-scale, Internet-derived data sets can be combined with existing traditional data from administrative procedures, surveys, and long-established government data sets such as the U.S. Census or Current Population Survey to create insights about behavioral patterns, management outcomes, operating anomalies, and so forth with a level of rapidity that has not previously been possible (Kitchin and McArdle 2016). Choi and Varian (2012) label this process "predicting the present." Administratively collected data always include a time lag of several months or even years. However, combining data from credit cards, Google, delivery services, or online shops can help government agencies forecast, for example, economic indicators such as unemployment (Llorente et al. 2015).

One advantage that scholars using data analytic approaches have is the ability to examine problems in real time—or at least more nearly in real time as capabilities grow.

Big data in public affairs, then, is *high-volume data that frequently combines highly structured administrative data actively collected by public sector organizations with continuously and automatically collected structured and unstructured real-time data that are often passively created by public and private entities through their Internet interactions.* Public sector organizations can make use of both administratively collected and unstructured, Internet-generated data to derive insights for their operations and public service delivery. Although public sector data science is increasingly comprehensive and granular in nature, it is currently biased in important ways, discussed next.

Methodologies, Theories, and Reinventing the Wheel

The promise of big data resides in the profusion of rich, prompt, granular data on behaviors and phenomena that were expensive and sometimes impossible to quantify in the past. For commercial enterprises, the exploitation of these data—so long as it stays within the legal frameworks in place—can often turbocharge the financial bottom line (McKinsey Global Institute 2011). However, for public organizations and scholars that wish to study them, the ambiguous, multifaceted, and contested "bottom line" of creating "public value" (Bryson, Crosby, and Bloomberg 2014; Moore 1995, 2014) generates a set of important questions and concerns. The issues outlined here must be carefully considered as practitioners increasingly rely on big data to inform their operations and as public affairs scholars use these data for their research.

Relying on Digital Exhaust

The term "digital exhaust" is apt for public affairs because much of what we think of as big data has not been captured with the purposes of public managers or researchers in mind.[1] In fact, most of it has been captured for purely technical reasons—for instance, automated rosters of log-ins to websites in order to track potential security breaches—with no exploitation expected beyond these mundane tasks. In other cases, data have been captured for commercial purposes that serve the needs of the collector but may only tangentially serve the needs of public affairs research or public operations.

Exploiting digital exhaust is tempting: the data sets can be extremely large, they can be quite comprehensive (populations rather than samples), and they are sometimes free (if the data provider is cooperative). However, recent experience with digital exhaust suggests that managers and scholars must proceed with caution. Lazer and colleagues (2014) remind

us that most digital exhaust does not rely on careful constructs that have been tested and found to be valid and reliable. Instead, these constructs are built (and often adapted) to fit a commercial or technical logic. Lazer et al. outline how changes to Google's search algorithm—updated for perfectly legitimate business reasons and further skewed by the endogenous behavior of Google search users and website owners—affected the predictions made by the Google Flu Trend (GFT) system.

There's Public and Then There Is Public

Digitization of data has helped make a wide range of "public" data available in cheap and easy-to-import formats. However, the notion of public has also shifted in the age of the Internet. "Public" records in the past were available but not particularly accessible. A prototypical example is information on housing tracts. It has long been possible with a visit to a municipal, county, or state office to learn a great deal about a given tract of land—size, shape, placement, improvements, transactions, and liens (mortgages, primarily). But it was not particularly easy to learn these facts because it required travel, copying, and fees, which meant that the transaction costs of acquiring this knowledge were nontrivial—but those with a strong motivation to know could know. Now, this information—and more—is increasingly available online; anyone who wants to access it can do so relatively freely, for the cost of Internet access and probably less than an hour of time. The Internet has greatly accelerated the possibilities for government to publish, access, or use information that has primarily lay hidden in paper archives but now is only a few clicks away.

In public affairs, citizens' unease with the perceived loss of privacy creates limits on the use of public data for both government operations and public affairs research.

In public affairs, citizens' unease with the perceived loss of privacy creates limits on the use of public data for both government operations and public affairs research. For government, use of such data can help improve services but can also undercut trust in government as citizens question the legitimacy of providing and accessing low-visibility data. For research, Institutional Review Boards have an uneven record regarding public data. On the one hand, personally identifiable information (PII) is usually tightly controlled through informed consent, data control plans, and strict data protection measures such as biometric access and "carryout" restrictions (see in particular DHHS 2009). On the other hand, protocols and procedures governing social media-generated data are less consistent. Is publicly available data created by social media users secondary data or primary data? In many cases, Institutional Review Boards allow researchers to use secondary data—that is, data from existing sources—without extensive review. Primary data—that is, data collected directly from research subjects—are much more closely regulated in terms of the methods for collection and storage. Simple, inexpensive web-scraping tools make it possible to ingest large data sets from public sites, opening new possibilities for research and evidence-based decision making—at the cost of greater visibility for previously undisclosed or virtually anonymous actions. As we highlight next, the very comprehensiveness of the data makes it possible to infer identity. While inferred identity could be a boon to scholars, it might also be a hindrance to public managers. Managers might like to use the data but are inhibited by U.S. privacy and information security laws that preclude collection of personally identifiable information from citizens visiting a government website (NIST 2010), which prevents them from realizing the potential benefits of this data source.

When 1 + 1 = I Know You

Privacy and digital anonymity are major concerns regarding use of digital data sets in the public sector. Using new matching techniques, it is already possible to link data from multiple sources—public and private—to develop far more comprehensive pictures of individuals and organizations than ever before. A 2014 Federal Trade Commission report on private data brokers found that the nine largest brokers held more than 3,000 "data segments" (that is, variables) on over 1.4 billion people worldwide, for a total of more than 700 billion data elements on individuals. With this, government is increasingly becoming a customer of data brokers to achieve legitimate public goals. For instance, in 2008, the U.S. Government Accountability Office reported that four large federal agencies "used personal information obtained from resellers for a variety of purposes, including performing criminal investigations, locating witnesses and fugitives, researching assets held by individuals of interest, and detecting prescription drug fraud" (GAO 2008, i). We also know that the federal government has ordered the release of telecommunications and credit card transactions data through the National Security Agency (Gorman, Perez, and Hook 2013). Bruce Schneier (2013) has characterized these practices as a back-door way around constitutional and statutory protections against unreasonable search and seizure of information by "ask[ing] corporate America for it." Further, transparency efforts by governments to release data make it easier to combine information from formerly disconnected government sources. Within this data-profuse landscape, researchers and public managers must be aware of the legal, ethical, and technical debates that are currently rearranging the regulatory frameworks that guide the use of data from public and private providers.

While government open data sources are de-identified, recent research has demonstrated just how easily multiple data sources may be leveraged to discover, for instance, one's Social Security number. Indeed, researchers from Carnegie Mellon University demonstrated in 2009 that common data elements found on most Facebook pages allow one to accurately predict the first five digits of a social security number (Acquisti and Gross 2009; Dannen 2009). With sufficient data, a great many pieces of identifying information could be reconstructed and behaviors could even be inferred (see, e.g., Target's ability to determine that someone is pregnant from buying habits in Duhigg 2012). While there is great promise in what we might learn from working with huge digital "dossiers" (Laudon 1986), there are also substantial concerns regarding the use of such comprehensive profiles of individuals for public affairs purposes.

Three-Quarters of a Digital Society

One presumption behind the enthusiasm for these data sets is that they are comprehensive. Yet there are substantial reasons to question the scope and representativeness of online data sources that are not

constructed and curated for research purposes. Scholars have often referred to these issues as problems of the "first-" and "second-" level digital divide. While some scholars have moved away from the term "digital divide" (DiMaggio et al. 2001) in favor of "digital inequality" (DiMaggio and Hargittai 2001; Hargittai and Hsieh 2013) or "digital differences" (Zickuhr and Smith 2012), the fact is that disparities in Internet access are still deeply entrenched. The Pew Research Center found in May 2013 that only 70 percent of U.S. households have high-speed broadband (Zickuhr and Smtih 2013) and that there are systematic differences between the broadband "haves" and "have nots."

The demographic factors most correlated with home broadband adoption continue to be educational attainment, age, and household income. Almost 9 in 10 college graduates have high-speed Internet at home, compared with just 37 percent of adults who have not completed high school. Similarly, adults under age 50 are more likely than older adults to have broadband at home, and those living in households earning at least $50,000 per year are more likely to have home broadband than those at lower income levels (Zickuhr and Smtih 2013).

While rural dwellers were long thought to be one source of differentiation, in fact, there are huge disparities across urban areas that correlate with poverty (from U.S. Census figures as summarized by Crow 2014). In addition, there are also huge disparities in the use of certain online social networking sites that are now used to glean insights and make generalizations regarding the whole population. For instance, only 17 percent of all Internet users belong to micromessaging services such as Twitter, but the tool has become one of the most important data sources for researchers (Duggan et al. 2015). Despite its unrepresentative nature, Twitter has become influential in initiating discussion of possible policy changes, such as those about police body cameras through the Black Lives Matter movement and the Transportation Security Administration's categorical screening mechanisms at airports. The insights and generalizations made from large data sets derived from Internet sources are thus often based on very explicit populations: Internet users who opt into using a specific social networking site. This practice might lead to false insights given that generalizations derived from study of Twitter users disregard the preferences and online behaviors of more than 80 percent of Internet users. Further, distractions based on the latest crisis and other immediacy pressures on social media could lead to goal displacement for the overall mission of governments (Lavertu 2016)—thus missing big questions about democracy and underrepresented communities (Kirlin 1996).

Research has also found substantial differences traced to educational attainment and socioeconomic class in the creation of online content such as blogs, websites, photos, videos, social network posts, chat rooms, and so on (Lutz and Hoffman 2014; Schradie 2011, 2015)—even among highly "wired" college students (Correa 2010). This disparity in content creation is sometimes called the "second-level" digital divide (Hargittai 2002). Early studies of content on the Internet suggest that content is disproportionately developed by people with more education and thus more income (Lutz and Hoffman 2014). Given that both income and education are correlated with race, ethnicity, and gender, content production is also likely to be biased across these three characteristics. Further, the poor, homeless, elderly, transient, and mentally and physically ill are all likely to be underrepresented in data constructed from Internet sites for the foreseeable future. Simply put, there is every reason to believe that digital exhaust from devices, automated text processing of online contributions, mining of digital profiles, and many other Internet sources will faithfully represent the biases found in Internet access, use, and content production patterns at the time the data are extracted.

Big Data, Little Theory

The editor of *Wired* magazine, Chris Anderson, spoke for many of our computational colleagues in 2008 when he declared "The End of Theory" and opined that petabyte-scale data

> forces us to view data mathematically first and establish a context for it later…. This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (2008, 71)

In its crudest form, exploitation of large-scale data sets is precisely the sort of data mining that every first-year social science PhD student is warned against in the compact form of the aphorism that "correlation is not causation." While Anderson's article represents the most fully distilled version of this line of thinking, the underlying ethos can be found in a wide range of publications extolling the benefits of analytics associated with large data sets. For example, a McKinsey Global Institute report (2011, see 27–31) catalogues a wide range of "big data techniques and technologies," most of which computationally induce relationships from large bodies of data—A/B testing, association rule learning, cluster analysis, machine learning, neural network analysis, genetic algorithms, and visualization techniques. Recent blog posts from government fellows at the IBM Center for the Business of Government also highlight uses of big data by U.S. federal agencies (Helms 2015a, 2015b, 2015c).

In the abstract, computational induction can be a valuable way to gain insights into new and unexplored problems. But unlike the social sciences, in which there is a tradition of using induction to inform theory development but also an expectation that additional data collection and analysis should be undertaken to confirm the resulting theory (see Behn 1995), data analysts often stop with correlation. The results have not always been edifying. In their discussion of the travails of Google Flu Trends, Lazer and colleagues provide a reference example of this approach gone awry:

> Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points. The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball. This should have been a warning

> sign that the big data were overfitting the small number of cases—a standard concern in data analysis. This ad hoc method of throwing out peculiar search terms failed when GFT completely missed the non-seasonal 2009 influenza A–H1N1pandemic. In short, the initial version of GFT was part flu detector, part winter detector. (Lazer et al. 2014, 1203)

The love affair with data science creates a danger that data, and not the issues raised by operations or constituencies, will drive public questions and analysis.

More professionally infuriating can be the instances in which computational approaches get the right answer—but one that social scientists have known about for 5, or 10, or 20 years. The combination of substantive ignorance of the subject matter plus excessive confidence in the method can lead to attributions of cause and effect that are situationally correct but globally faulty. As the GFT example demonstrates, reliance solely on analysis of Internet data can lead to serious (here, quite literally deadly) consequences when theory and context are ignored. The "parable of GFT," to use Lazer's phrase, suggests that a stronger foundation for big data initiatives in public affairs is one that engages with 150 years of public affairs theory and analysis rather than setting it aside. Even analysts at Google and Facebook warn against using their companies' vast lakes of data without appropriate "small data" to provide context and nuance (Peysakhaoich and Stephens-Davidowitz 2015). With that said, unless public affairs as a field is able to demonstrate its ability to engage with data analytics productively and responsibly, others will supplant us.

Implications

The big data era is upon us, and there is momentum to leverage that data to provide new insights for the public good. But these new opportunities have implications for both practitioners and policy makers.

Public managers and policy makers operate in the context of many constraints: limited budgets, multiple constituencies, and short time frames in which to make and implement decisions. Big data accumulates quickly and seemingly exponentially; it can quickly overwhelm an analyst. Public managers will need the capability to (1) manage and process large accumulations of unstructured, semistructured, and structured data; (2) analyze that data into meaningful insights for public operations; and (3) interpret that data in ways that support evidence-based decision making. We use the term "capability" here advisedly, as public managers will likely use a mix of staff, contractors, and personal resources to manage, analyze, and interpret large-scale data sets, be they administrative or Internet based. These new capabilities will require different resources than the typical public bureaucracy has now—especially in smaller jurisdictions.

In a recent article, Lavertu (2016) illustrates just how insidious data analytics can be. He points out that information derived from these data sets gives informed citizens a false feeling of precision with respect to how well we can actually measure some of the important outcomes sought by public programs. Few people examine carefully how those measurement decisions were made, how the data were generated, and what the strengths and weaknesses of the approach are more generally. It further creates the conundrum of providing information at uneven levels of precision, where often the more granular data are (perhaps subconsciously) favored inappropriately only because they are more granular. All of this can divert attention from the broader goals.

The love affair with data science creates a danger that data, and not the issues raised by operations or constituencies, will drive public questions and analysis. While the data may be able to reveal problems previously obscured, there should be a weighting between the two approaches that is appropriate for a given jurisdiction. Administrators need to have—or to have available—the analytical capabilities required to perform just-in-time analyses of large-scale data sets. Additionally, policy makers must understand the benefits and weaknesses of relying on large-scale data sets and the techniques used to analyze them. These data and analysis of them exacerbate known problems with current approaches to empirically based policy making (Brownson, Fielding, and Maylahn 2009).

With respect to public affairs research, one of the most substantial and sustained critiques of scholarship is that the insights we derive are too late or not of immediate relevance to policy makers (Isett, Head, and VanLandingham 2016; Jewell and Bero 2008). Just-in-time analysis or other data techniques can reduce the lag between emergence of public problems and formulation of research-informed solutions. Currently, public affairs research lags public practice because designing and implementing data collection processes is time and resource intensive. In a big data world, the data are there, so the scholar can delve into the data after devising only an analytical strategy. Data availability will shorten the lag between identified problems and results worth using. But, of course, this invites the questions about whether the data available are appropriate for answering the question or merely convenient, whether this is the kind of work that can and will dominate future public debates, and whether these approaches and capabilities should be developed instead of other techniques that have longer lags.

While there is a lot of talk about big data and its promise, to what extent will things really change? After all, most of the studies on the Government Performance and Results Act (GPRA) of 1993—the largest performance management reform in recent history—did not yield much actual change (Moynihan 2005; Moynihan and Ingraham 2003). In fact, almost all studies of performance reform suggest that these efforts have very little impact (Gerrish 2016). Yet there is a glimmer of hope specific to big data, as recent analysis of the GPRA Modernization Act of 2010 (Moynihan and Kroll 2016) suggests that the Modernization Act routines did improve performance information use. Nevertheless, while big data has the potential to enhance the operations of government, it can also be a threat to individual managers' self-efficacy through changes in the portfolio of skills that are valued in public agencies (Choi and Varian 2012; Lavertu 2016; Wright, Christensen, and Isett 2013).

Conclusion

Big data is here, and public administrators need to grapple with what this means for the field. While the field is trying to

determine how to leverage data science for the public good, we watch computer scientists rediscover our findings and claim them for their own. The confirmatory findings circulating among data analytic specialists is both gratifying and frustrating. It is gratifying because our theories and models are being borne out by gigabytes and terabytes of data demonstrating the relevance and rigor of our research. But it is also frustrating because the new computationalists are reinventing our research out of ignorance. Regardless of the tensions between social and computer science research, it is clear that there are important considerations for large data sets and data analytics in the public sector.

From an operational perspective, we raise several questions. First, there are ethical considerations. Who can use public data and for what purposes? Second, there are privacy considerations. PII can be scraped, cobbled together, or bought outright to create a fairly specific profile of individuals. How much data should government be allowed to harvest and hold from citizens, and for what ends? Third, there are secrecy and security concerns. How will PII be protected to ensure that it is not used for evil or fall into the hands of evildoers? Fourth, there are effectiveness concerns. How much data can government actually use or analyze effectively? Do we trust public entities to do it well? Finally, there are feasibility and efficacy concerns. What are the legitimate and worthwhile uses of digital exhaust? To what extent does digital exhaust provide an unbiased source of data for public decision making and research? How much can and should be leveraged for the public good?

From a scholarship perspective, there are different issues related to the "big questions" in public affairs with which we need to grapple. Perhaps the most pressing issue concerns representation and democracy. How does the big data movement affect how citizens' voices are heard and acted upon? This is a fundamental question of public affairs. Data analytics raises the specter of voices being filtered through data rather than coming directly from citizens. While data can undoubtedly help us identify problems, whose problems are they? And are they the most salient problems for a particular community? A related big question is about the workforce in public affairs. If the field becomes highly data driven, with a skew toward data analytics, then our workforce will inevitably move toward technocratic operations rather than a mix of the technical and the humanist. This is an important balance that ought to be deliberated rather than allowing the field to organically drift to one option or the other.

All of the issues raised in this article are important for moving forward in the public sector with big data. Even if they are flawed, large-scale data sets and the tools used to analyze them provide windows into populations and behaviors that are otherwise too expensive or difficult to collect. Data that are derived from the Internet and administrative sources but creatively connected together can provide immediate, operationally relevant insights that small data techniques simply cannot. And large-scale data sets from Internet sources will continue to improve if content and participation becomes more pluralistic and inclusive. Nevertheless, large-scale data sets, whatever their sources, are not a replacement for theory and small data techniques. Theory is still an invaluable guide to data analysis (small and big), and small data built to answer specific questions can still provide the most precise answers. As a field, we will be challenged educationally and intellectually to better understand the evolving limits and opportunities to using big data and to identify the set of indispensable data analytic skills we must impart to our professional and research students. We will also be challenged with respect to how to marry our theory and insights to the tools our computational colleagues are regularly churning out. We fully admit to not having all the answers, but we believe that posing these "big questions" is the first step toward finding solutions that move forward the field and practice of public management.

Note

1. A brief search of the Internet did not turn up a progenitor for this term. One of the earliest references was in a blog post by Steven Mandzik from December 15, 2007, http://stevenmandzik.com/web-20/digital-exhaust/ (accessed July 28, 2016).

References

Acquisti, Alessandro, and Ralph Gross. 2009. Predicting Social Security Number from Public Data. *Proceedings of the National Academy of Sciences of the United States of America* 106(27): 10975–80.

Anderson, Chris. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired,* June 23. http://www.wired.com/2008/06/pb-theory/ [accessed July 25, 2016].

Behn, Robert D. 1995. The Big Questions in Public Management. *Public Administration Review* 55(4): 313–24.

Boyd, Danah, and Kate Crawford. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15(5): 662–79.

Brownson, Ross C., Jonathan E. Fielding, and Christopher M. Maylahn. 2009. Evidence–Based Public Health: A Fundamental Concept for Public Health Practice. *Annual Review of Public Health* 30: 175–201.

Bryant, Randal E., Randy H. Katz, and Edward D. Lazowska. 2008. Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science, and Society. White paper prepared for the Computing Community Consortium Committee of the Computing Research Association. http://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf [accessed July 25, 2016].

Bryson, John M., Barbara C. Crosby, and Laura Bloomberg. 2014. Introduction: Public Value Governance: Moving Beyond Traditional Public Administration and the New Public Management. *Public Administration Review* 74(4): 445–56.

Chen, Hsinchun, Roger H. L. Chiang, and Veda C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36(4): 1165–88.

Choi, Hyunyoung, and Hal Varian. 2012. Predicting the Present with Google Trends. Supplement 1. *Economic Record* 88: 2–9.

Clark, William Roberts, and Matt Golder. 2015. Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science? *PS: Political Science and Politics* 48(1): 65–70.

Correa, Teresa. 2010. The Participation Divide among "Online Experts": Experience, Skills, and Psychological Factors as Predictors of College Students' Web Content Creation. *Journal of Computer-Mediated Communication* 16(1): 71–92.

Crow, David. 2014. Digital Divide Exacerbates U.S. Inequality. *Financial Times,* October 28. http://www.ft.com/cms/s/2/b75d095a-5d76-11e4-9753-00144feabdc0.html [accessed July 25, 2016].

Dannen, Chris. 2009. On Facebook? New Algorithm Can Guess Your SSN. *Fast Company,* July 7. http://www.fastcompany.com/1305136/facebook-new-algorithm-can-guess-your-ssn [accessed July 25, 2016].

Denning, Peter J. 1990. The Science of Computing: Saving All the Bits. *American Scientist* 78: 402–5.

DiMaggio, Paul J., and Eszter Hargittai. 2001. From "Digital Divide" to "Digital Inequality": Studying Internet Use as Penetration Increases. Working Paper no. 15, Princeton University, Center for Arts and Cultural Policy. https://www.princeton.edu/~artspol/workpap15.html [accessed July 25, 2016].

DiMaggio, Paul J., Hargittai Eszter, W. Russell Newman, and John P. Robinson. 2001. Social Implication of the Internet. *Annual Review of Sociology* 27: 307–36.

Doctorow, Cory. 2008. Big Data: Welcome to the Petacentre. *Nature* 455: 16–21.

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Leinhart, and Mary Madden. 2015. *Social Media Update 2014. Pew Research Center Internet Science Technology*, January 9. http://www.pewinternet.org/2015/01/09/social-media-update-2014/ [accessed July 25, 2016].

Duhigg, Charles. 2012. How Companies Learn Your Secrets. *New York Times Magazine,* February 16. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html [accessed July 28, 2016].

Eagle, Nathan, Alex Pentland, and David Lazer. 2009. Inferring Friendship Network Structure by Using Mobile Phone Data. *Proceedings of the National Academy of Sciences* 106(36): 15274–78.

Federal Trade Commission. 2014. *Data Brokers:* A Call for Transparency and Accountability. https://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014 [accessed July 25, 2016].

Frankel, Felice, and Rosalind Reid. 2008. Distilling Meaning from Data. *Nature* 455: 30.

George, Gerard, Martine R. Haas, and Alex Pentland. 2014. Big Data and Management. *Academy of Management Journal* 57(2): 321–26.

Gerrish, Edwin. 2016. The Impact of Performance Management on Performance in Public Organizations: A Meta-Analysis. *Public Administration Review* 76(1): 48–66.

Gorman, Siobhan, Evan Perez, and Janet Hook. 2013. U.S. Collects Vast Data Trove. *Wall Street Journal,* June 7. http://www.wsj.com/articles/SB10001424127887324299104578529112289298922 [accessed July 25, 2016].

Hargittai, Eszter. 2002. Second-Level Digital Divide: Difference in People's Online Skills. *First Monday* 7(4). http://firstmonday.org/article/view/942/864 [accessed July 25, 2016].

Hargittai, Eszter, and Yu-li Patrick Hsieh. 2013. Digital Inequality. In *Oxford Handbook of Internet Studies*, edited by William H. Dutton, 129–50. Oxford, UK: Oxford University Press.

Helms, Josh. 2015a. Challenges in Adopting a Big Data Strategy (Part 1 of 2). *Business of Government Blog,* February 11. http://www.businessofgovernment.org/blog/business-government/challenges-adopting-big-data-strategy-part-1-2 [accessed July 25, 2016].

———. 2015b. Challenges in Adopting a Big Data Strategy (Part 2 of 2). *Business of Government Blog,* February 18. http://www.businessofgovernment.org/blog/business-government/challenges-adopting-big-data-strategy-part-2-2 [accessed July 25, 2016].

Helms, Josh. 2015c. Five Examples of How Federal Agencies Use Big Data. *Business of Government Blog,* February 25. http://www.businessofgovernment.org/blog/business-government/five-examples-how-federal-agencies-use-big-data [accessed July 25, 2016].

Isett, Kim R., Brian W. Head, and Gary VanLandingham. 2016. Caveat Emptor: What Do We Know about Public Administration Evidence and How Do We Know It? *Public Administration Review* 76(1): 20–23.

Janssen, Marijn, and Jeroen van den Hoven. 2015. Big and Open Linked Data (BOLD) in Government: A Challenge to Transparency and Privacy? *Government Information Quarterly* 32(4): 363–68.

Jewell, Christopher J., and Lisa A. Bero. 2008. Developing Good Taste in Evidence: Facilitators of and Hindrances to Evidence-Informed Health Policymaking in State Government. *Milbank Quarterly* 86(2): 177–208.

Johnston, Erik. 2015. *Governance in the Information Era: Theory and Practice of Policy Informatics*. New York: Routledge.

Kirlin, John J. 1996. The Big Questions of Public Administration in a Democracy. *Public Administration Review* 56(5): 416–23.

Kitchin, Rob, and Gavin McArdle. 2016. What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society.* http://bds.sagepub.com/content/3/1/2053951716631130 [accessed July 25, 2016].

Laudon, Kenneth C. 1986. *Dossier Society: Value Choices in the Design of National Information Systems*. New York: Columbia University Press.

Lavertu, Stéphane. 2016. We All Need Help: "Big Data" and the Mismeasure of Public Administration. *Public Administration Review* 76(6): 864–72.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176): 1203–5.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, et al. 2009. Computational Social Science. *Science* 323(5915): 721.

Llorente, Alejandro, Manuel Garcia-Herranz, Manuel Cebrian, and Esteban Moro. 2015. Social Media Fingerprints of Unemployment. *PLoS One* 10(5): e0128692.

Lutz, Christoph, and Christian Pieter Hoffman. 2014. Towards a Broader Understanding of the Participation Divide(s). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2436154 [accessed July 25, 2016].

Lynch, Clifford. 2008. Big Data: How Do Your Data Grow? *Nature* 455: 28–29.

McAfee, Andrew, and Erik Brynjolfsson. 2012. Big Data. The Management Revolution. *Harvard Business Review* 90(10): 61–67.

McKinsey Global Institute. 2011. Big Data: The Next Frontier of Innovation, Competition, and Productivity. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation [accessed July 28, 2016].

Meier, Kenneth J., and Laurence J. O'Toole, Jr. 2005. Managerial Networking Issues of Measurement and Research Design. *Administration & Society* 37(5): 523–41.

Moore, Mark H. 1995. *Creating Public Value: Strategic Management in Government*. Cambridge, MA: Harvard University Press.

——— 2014. Public Value Accounting: Establishing the Philosophical Basis. *Public Administration Review* 74(4): 465–77.

Moynihan, Donald P. 2005. Why and How Do State Governments Adopt and Implement "Managing for Results" Reforms? *Journal of Public Administration Research and Theory* 15(2): 219–43.

Moynihan, Donald P., and Patricia W. Ingraham. 2003. Look for the Silver Lining: When Performance-Based Accountability Systems Work. *Journal of Public Administration Research and Theory* 13(4): 469–90.

Moynihan, Donald P., and Alexander Kroll. 2016. Performance Management Routines That Work? An Early Assessment of the GPRA Modernization Act. *Public Administration Review* 76(2): 314–23.

National Institute of Standards and Technology (NIST). 2010. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII): Recommendations of the National Institute of Standards and Technology. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=904990 [accessed July 28, 2016].

Onnela, J. P., J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, and J. Kertész. 2007. Structure and Tie Strengths in Mobile Communication Networks. *Proceedings of the National Academy of Sciences* 104(18): 7332–36.

Peysakhaoich, Alex, and Seth Stephens-Davidowitz. 2015. How Not to Drown in Numbers. *New York Times,* May 2. http://www.nytimes.com/2015/05/03/opinion/sunday/how-not-to-drown-in-numbers.html [accessed July 28, 2016].

Pirog, Maureen A. 2014. Data Will Drive Innovation in Public Policy and Management Research. *Journal of Policy Analysis and Management* 33(2): 537–43.

Schneier, Bruce. 2013. Do You Want the Government Buying Your Data from Corporations? *The Atlantic,* April 30. http://www.theatlantic.com/technology/

archive/2013/04/do-you-want-the-government-buying-your-data-from-corporations/275431/ [accessed July 25, 2016].

Schradie, Jen. 2011. The Digital Production Gap: The Digital Divide and Web 2.0 Collide. *Poetics* 39(2): 145–68.

———. 2015. The Gendered Digital Production Gap: Inequalities of Affluence. In *Communication and Information Technologies Annual,* edited by Laura Robinson, Sheila R. Cotton, and Jeremy Schulz, 185–213. Bingley, UK: Emerald Group.

U.S. Department of Health and Human Services (DHHS). 2009. Basic HHS Policy for Protection of Human Research Subjects—Code of Federal Regulations. Department of Health and Human Services. http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html [accessed July 25, 2016].

U.S. Government Accountability Office (GAO). 2008. Privacy: Government Use of Data from Information Resellers Could Include Better Protections-Statement of Linda D. Koontz, Director, Information Management Issues. Washington, DC: U.S. Government Printing Office. GAO-08-543T. http://www.gao.gov/htext/d08543t.html [accessed July 25, 2016].

White House. 2014. Big Data: Seizing Opportunities, Preserving Values. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf [accessed July 25, 2016].

Wright, Bradley E., Robert Christensen, and Kimberley R. Isett. 2013. Motivated to Adapt? The Role of Public Service Motivation as Employees Face Organizational Change. *Public Administration Review* 73(5): 738–46.

Zickuhr, Kathryn, and Aaron Smith. 2012. Digital Differences. *Pew Research Center*, April 13. http://www.pewinternet.org/files/old-media//Files/Reports/2012/PIP_Digital_differences_041312.pdf [accessed July 25, 2016].

———. 2013. Home Broadband 2013. Pew Research Center, August 26. http://www.pewinternet.org/2013/08/26/home-broadband-2013/ [accessed July 25, 2016].