© Academy of Management Journal
2014, Vol. 57, No. 2, 321–326.
http://dx.doi.org/10.5465/amj.2014.4002

FROM THE EDITORS

BIG DATA AND MANAGEMENT

Editor's note: This editorial launches a series written by editors and co-authored with a senior executive, thought leader, or scholar from a different field to explore new content areas and grand challenges with the goal of expanding the scope, interestingness, and relevance of the work presented in the Academy of Management Journal. *The principle is to use the editorial notes as "stage setters" for further work and to open up fresh, new areas of inquiry for management research. GG*

Big data is everywhere. In recent years, there has been an increasing emphasis on big data, business analytics, and "smart" living and work environments. Though these conversations are predominantly practice driven, organizations are exploring how large-volume data can usefully be deployed to create and capture value for individuals, businesses, communities, and governments (McKinsey Global Institute, 2011). Whether it is machine learning and web analytics to predict individual action, consumer choice, search behavior, traffic patterns, or disease outbreaks, big data is fast becoming a tool that not only analyzes patterns, but can also provide the predictive likelihood of an event.

Organizations have jumped on this bandwagon of using ever-increasing volumes of data, often in tera- or petabytes' worth of storage capacity, to better predict outcomes with greater precision. For example, the United Nations' Global Pulse is an initiative that uses new digital data sources, such as mobile calls or mobile payments, with real-time data analytics and data mining to assist in development efforts and understanding emerging vulnerabilities across developing countries. Though "big data" has now become commonplace as a business term, there is very little published management scholarship that tackles the challenges of using such tools—or, better yet, that explores the promise and opportunities for new theories and practices that big data might bring about. In this editorial, we explore some of its conceptual foundations as well as possible avenues for future research and application in management and organizational scholarship.

WHAT IS "BIG DATA"?

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 or business transactions such as sales queries and purchase transactions. In addition, genomics, health care, engineering, operations management, the industrial Internet, and finance all add to big data pervasiveness. These data require the use of powerful computational techniques to unveil trends and patterns within and between these extremely large socioeconomic datasets. New insights gleaned from such data-value extraction can meaningfully complement official statistics, surveys, and archival data sources that remain largely static, adding depth and insight from collective experiences—and doing so in real time, thereby narrowing both information and time gaps.

Perhaps the misnomer is in the "bigness" of big data, which invariably attracts researchers' attention to the size of the dataset. Among practitioners, there is emergent discussion that "big" is no longer the defining parameter, but, rather, how "smart" it is—that is, the insights that the volume of data can reasonably provide. For us, the defining parameter of big data is the fine-grained nature of the data itself, thereby shifting the focus away from the number of participants to the granular information about the individual. For example, a participant in a Formula 1 car race generates 20 gigabytes of data from the 150 sensors on the car that can help analyze the technical performance of its components, but also the driver reactions, pit stop delays, and communication between crew and driver that contribute to overall performance (Munford, 2014). The emphasis thus moves away from outcomes (win/lose race), to instead focus on each proximal, contributory element for success or failure mapped for every second during the race. Similarly, one could analyze the social networks and social engagement behaviors of individuals by mapping mobility patterns onto physical layouts of workspaces using sensors, or the frequency of meeting room usage using remote sensors that track entry and exit patterns, which could provide information on communication and coordination needs based on project complexity and approaching deadlines. These micro data provide a richness of individual behaviors and actions that have not yet been fully tapped in management research. Whether it is "big" or "smart" data, the use of large-scale data to predict human behavior is gaining currency in business and government policy practice, as well as in scientific domains where the physical and social sciences converge (recently referred to as "social physics") (Pentland, 2014).

Sources of Big Data

Big data is also a wrapper for different types of granular data. Below, we list five key sources of high volume data: (1) public data, (2) private data, (3) data exhaust, (4) community data, and (5) self-quantification data.

"Public data" are data typically held by governments, governmental organizations, and local communities that can potentially be harnessed for wide-ranging business and management applications. Examples of such data include those concerning transportation, energy use, and health care that are accessed under certain restrictions in order to guard individual privacy. "Private data" are data held by private firms, non-profit organizations, and individuals that reflect private information that cannot readily be imputed from public sources. For example, private data include consumer transactions, radio-frequency identification tags used by organizational supply chains, movement of company goods and resources, website browsing, and mobile phone usage, among several others.

"Data exhaust" refers to ambient data that are passively collected, non-core data with limited or zero value to the original data-collection partner. These data were collected for a different purpose, but can be recombined with other data sources to create new sources of value. When individuals adopt and use new technologies (e.g., mobile phones), they generate ambient data as by-products of their everyday activities. Individuals may also be passively emitting information as they go about their daily lives (e.g., when they make purchases, even at informal markets; when they access basic health care; or when they interact with others). Another source of data exhaust is information-seeking behavior, which can be used to infer people's needs, desires, or intentions. This includes Internet searches, telephone hotlines, or other types of private call centers.

"Community data" is a distillation of unstructured data—especially text—into dynamic networks that capture social trends. Typical community data include consumer reviews on products, voting buttons (such as, "I find this review useful"), and Twitter feeds, among many others. These community data can then be distilled for meaning to infer patterns in social structure (e.g., Kennedy, 2008). "Self-quantification data" are types of data that are revealed by the individual through quantifying personal actions and behaviors. For example, a common form of self-quantification data is that obtained through the wristbands that monitor exercise and movement, data which are then uploaded to a mobile phone application and can then be tracked and aggregated. In psychology, individuals have "stated preferences" of what they would like to do versus "revealed preferences," wherein the preference for an action or behavior is inferred. For example, an individual might buy energy-efficient lightbulbs with the goal of saving electricity, but, instead, keep the lights on longer because they are now using less energy. Such self-quantification data helps bridge the connection between psychology and behavior. Social science scholars from diverse areas, such as psychology, marketing, or public policy, could benefit from stated and implicit preference data for use in their research.

Data Sharing, Privacy, and Ethics

In current information technology infrastructures, the provision of services such as network connectivity is usually associated with a Service Level Agreement (SLA) defining the nature and quality of the service to be provided. Such SLAs are important to limit liability, to enable better provisioning of the operational infrastructure for the provider, and to provide a framework for differential pricing. The exponential expansion of network connectivity and web services was, in large part, due to significant technological advances in the automation of SLA enforcement, in terms of monitoring and verification of compliance with the contract. In contrast, the realm of big data-sharing agreements remains informal, poorly structured, manually enforced, and linked to isolated transactions (Koutroumpis & Leiponen, 2013). This acts as a significant barrier to the market in data—especially for social science and management research, which cannot access these private data for integration with other public sources.

Data sharing agreements need to be linked into the mechanisms for data protection and privacy, including anonymization for open data, access control, rights management, and data usage control. Issues such as imputed identity, where individual identity can be inferred through data triangulation from multiple sources, will need to be carefully considered and explicitly acknowledged and permitted. Management scholars will be invited to embed themselves into social issues based on defining research questions that integrate data sharing and privacy as part of their research methodology. Doing so will likely allow us to refine the model for data sharing and data rights, which could be universally beneficial and define big data collaborations in the future.

ANALYZING BIG DATA

Equally relevant as the sources of data are the methodologies to analyze them and the standards of evidence that would be acceptable to management scholars for their publication. As with any nascent science, there is likely to be a trade-off between theoretical and empirical contribution, and the rigor with which data are analyzed. Perhaps, with big data, we are liable to initially be confounded by the standard of evidence that should be expected. The typical statistical approach of relying on p values to establish the significance of a finding is unlikely to be effective because the immense volume of data means that almost everything is significant. Using our typical statistical tools to analyze big data, it is very easy to get false correlations. However, this doesn't necessarily mean that we should be moving toward more and more complex and sophisticated econometric techniques to deal with this problem; indeed, such a response poses a substantial danger of over-fitting the data. Instead, basic Bayesian statistics and stepwise regression methods may well be appropriate approaches. Beyond these familiar approaches, there is a range of specialized techniques for analyzing big data, each of which is important for those entering this field to understand, though beyond the scope of this editorial. These techniques draw from several disciplines, including statistics, computer science, applied mathematics, and economics. They include (but are not limited to) A/B testing, cluster analysis, data fusion and integration, data mining, genetic algorithms, machine learning, natural language processing, neural networks, network analysis, signal processing, spatial analysis, simulation, time series analysis, and visualization (McKinsey Global Institute, 2011).

The challenge, though, is to shift away from focusing on p values to focusing, rather, on effect sizes and variance explained. With further empirical work, perhaps scholars can develop and converge on rough heuristics; for example, an R^2 of more than 0.3 could suggest that closer scrutiny of the pattern of relationships is warranted. Another pitfall of big data—again, amplified by our commonly used statistical techniques—lies in focusing too much on aggregates or averages and too little on outliers. In many situations, averages are very important, and often revealing about how people tend to behave under particular conditions. But, in the vastness of a big data universe, the outliers can be even more interesting: critical innovations, trends, disruptions, or revolutions may well be happening outside the average tendencies, yet still involve enough people to have dramatic effects over time. The fine-grained nature of big data offers opportunities to identify these sources of change—be they business innovations, social trends, economic crises, or political upheavals—as they gather steam.

Once promising leads have been identified, the next challenge of analyzing big data is to then move beyond identifying correlational patterns to exploring causality. Given the unstructured nature of most big data, causality is not built into their design and the patterns observed are often open to a wide range of possible causal explanations. There are two main ways to approach this issue of causality. The first is to recognize the central importance of theory. An intuition about the causal processes that generated the data can be used to guide the development of theoretical arguments, grounded in prior research and pushing beyond it. The second, complementary way is to then test these theoretical arguments in subsequent research—ideally, through field experiments. Of course, laboratory experiments offer the advantage of greater control, but they usually focus on a very limited number of variables, and the nature of big data research is that there may be many factors driving the observed correlational patterns. In a field experiment, a wider net can be cast, as a richer set of data about behaviors and beliefs can be collected, and over an extended period of time. For scholars as well as managers with an interest in action research, there are alluring opportunities here to engage in "management engineering" that goes beyond more typical management research by bringing theory and practice together with much faster cycle times between the identification of a promising theoretical insight and the testing of that insight with a well-designed intervention that can help to both advance management knowledge and address pressing practical questions.

Ultimately, the promise and the goal of strong management research built on big data should be not only to identify correlations and establish plausible causality, but, ultimately, to reach *consilience*—that is, convergence of evidence from multiple, independent, and unrelated sources, leading to strong conclusions (Wilson, 1998). Big data offers exciting new prospects for achieving such consilience due to its unprecedented volume, micro-level detail, and multifaceted richness. The vast majority of current management research relies on painstaking collection of low numbers of measures that cover a short duration of time (or, possibly, in the case of more historically based research, a longer duration but comprised of larger periods, such as years). In contrast, big data offers voluminous quantities of data over multiple periods (whether seconds, minutes, hours, days, months, or years).

While some big data datasets are unidimensional or single channel, focusing, for example, on a particular transaction or communication behavior and relying on single-channel interactions (e.g., via phone or email), there are increasingly opportunities to collect and analyze multidimensional datasets that offer insight into constellations of behaviors, often through a variety of channels (e.g., call center customer interactions that switch between voice, web, chat, mobile, video, etc). For management researchers, the result of such richness is that there are unprecedented opportunities to notice potentially important variables that previous studies might have failed to consider at all, due to their necessarily more focused nature. And, once such variables capture a researcher's attention, the relationships between them can be explored and the contextual conditions under which these relationships may or may not hold can be examined.

BIG DATA IN MANAGEMENT RESEARCH

Our intent in this editorial is to encourage fresh, new areas of scholarly inquiry—it is not to provide a systematic review of big data applications; neither do we pretend to provide a definitive guide for future research. Instead, our goal is to trigger broader discussions of big data in society and its implications for management research. The constantly changing environment in the digital economy has challenged traditional economic and business concepts. Huge volumes of user-generated data are transferred and analyzed within and across different sectors, gradually increasing the markets' dependency on precise and timely information services. A mere Tweet from a trusted source can cause losses or profits of billions of dollars and a chain reaction in the press, social networks, and blogs. This situation makes information goods even more difficult to value, as they have a catalytic impact on real-time decision making. Meanwhile, entrepreneurs and innovators have taken aggregate open and public data as well as community, self-quantification and exhaust data to create new products and services that have the power to transform industries. In private and public spheres, big data sourced from mobile technologies and banking services, such as digital/mobile money, when combined with existing "low-tech" services, such as water or electricity, can transform societies and communities. There is little doubt that, over the next decade, big data will change the landscape of social and economic policy and research.

What is unclear is how these "new models" for mixing and matching these products, services and data come about and evolve into a sustainable social and economic model. Categorizing big data, assessing its quality, and identifying its impact is radically new in social sciences, especially in management and organizational research. The rate and scale of content generation multiplies its impact and diminishes the time to respond. Consequently, management scholars will need to unpack how ubiquitous data can generate new sources of value, as well as the routes through which such value is manifest (mechanisms of value creation) and how this value is apportioned among the parties and data contributors, entrepreneurs, businesses, industries, and governments through new business models and new governance tools, such as contracts and licenses (mechanisms of value capture).

Empirical research in management often infers relationships; for example, two companies might be competing in the same market, have complementary products, collaborate in production or R&D, or be linked through supplier–customer relationships, or they might be close to each other in geographic, technology, or some other space that might facilitate knowledge spill-overs between them.

Detailed data on these relationships are typically unavailable in firm-level datasets that allow representative statistical inference. However, information on such relationships is often available in unstructured textual form, such as in news articles or company blogs on the web. IBM estimates that as much as 80% of this relationship information is unstructured "content" of various communications through email, texts, and videos—and they reckon unstructured content data is growing at twice the rate of conventional structured databases. To address such data, content analytics is emerging as a commercial evolution of what academics call "content analysis," or the analysis of text and other kinds of communication for the purposes of identifying robust patterns.

There are additional uses of big data that have broader implications for communities and societies, but which managers would find useful. For example, disease spreads, commuting patterns, or emotions and moods of communities, which can all be accessed through live Twitter feeds or Facebook postings, could affect organizational responses, products and services, and their strategies. Patterns in social media are being used to gleam information on the creation of new markets and product categories. Many companies now use digital intervention labs that track social media on a real-time basis around the world, thereby creating longitudinal data structures of millions of posts, Tweets, or reviews. Any deviations from normal patterns that invoke their brand or products are immediately flagged for action to provide rapid responses to consumer reactions, shape new product introductions, and create new markets.

The continuous, ubiquitous nature of the data means that scholars have a wealth of new opportunities to focus on the microfoundations of organizational strategies or behaviors; for instance, we can examine the dynamics of how business processes and opportunities evolve on a minute-to-minute, day-to-day basis, rather than being constrained to assessing snapshots such as quarterly inputs and outcomes or sales cycle trends. Consider the famous example of the Hubble space telescope having the wrong optics installed because one group assumed metric measurements and another imperial measurements, or the example of the Airbus 380 in which the wiring harness built in Germany and Spain did not fit the airframe built in Britain and France because the standards adopted were different. Current practice would be to review procedures and suggest more checkpoints; that is, a relatively static measurement and control of organizational actions. Instead, we could use big data to check what sort of communication patterns are required to avoid such disasters, and where we might discover that the lack of face-to-face communication at the "alpha test" stage was the critical variable, we might then suggest establishment of a real-time data-monitoring mechanism to ensure that face-to-face communication happened at all the necessary "alpha test" junctures.

Big data can also be a potent tool for analysis of individual or team behavior, using sensors or badges to track individuals as they work together, move around their workspace, or spend time interacting with others or allocated to specific tasks. While early management research codified diaries and time-management techniques of CEOs, evolving practices—using big data—can allow us to study entire organizations and workgroups in near-real time to predict individual and group behaviors, team social dynamics, coordination challenges, and performance outcomes. Scholars could examine questions around the differences between stated versus revealed preferences by tracking data on purchasing, mobile applications, and social media engagement and consumption, to state but a few examples. Social network studies could also use big data to examine the dynamics of formal and informal networks as they form and evolve, as well as their impact on individual, network, and organizational behaviors. Such granular, high-volume data can tell us more about workplace practices and behaviors than our current data-collection methods allow—and have the potential to transform management theory and practice.

Gerard George
Imperial College London

Martine R. Haas
University of Pennsylvania

Alex Pentland
MIT

REFERENCES

Kennedy, M. T. 2008. Getting counted: Markets, media, and reality. ***American Sociological Review,*** 73: 270–295.

Koutroumpis, P., & Leiponen, A. 2013. Understanding the value of (big) data. In ***Proceedings of 2013 IEEE international conference on big data.*** 38–42. Silicon Valley, CA, October 6–9, 2013. Los Alamitos, CA: IEEE Computer Society Press.

McKinsey Global Institute. 2011. ***Big data: The next fron-***

tier for innovation, competition, and productivity. June 2011. Lexington, KY: McKinsey & Company.

Munford, M. 2014. Rule changes and big data revolutionise Caterham F1 chances. ***The Telegraph,*** Technology Section, 23 February 2014. Available from http://www.telegraph.co.uk/technology/technology-topics/10654658/Rule-changes-and-big-data-revolutionise-Caterham-F1-chances.html.

Pentland, A. 2014. ***Social Physics.*** New York, NY: Penguin.

Wilson, E. O. 1998. ***Consilience: The unity of knowledge.*** New York, NY: Knopf.

Gerry George is professor of innovation and entrepreneurship at Imperial College London and serves as deputy dean of the Business School. He is the editor of the *Academy of Management Journal*.

Martine Haas is associate professor of management at the Wharton School at the University of Pennsylvania. She is an associate editor of the *Academy of Management Journal*, covering the topics of knowledge management, multinationals, and organization theory.

Alex "Sandy" Pentland is Toshiba professor of media, arts, and sciences and director of the MIT Media Lab Entrepreneurship Program at MIT. He is a pioneer in organizational engineering, mobile information systems, and computational social science. His research focus is on harnessing information flows and incentives within social networks, the big data revolution, and converting this technology into real-world ventures. He is the World Economic Forum's lead academic for its big data and personal data initiatives. He is among the most-cited computer scientists in the world, and, in 1997, *Newsweek* magazine named him one of the 100 Americans likely to shape this century. His book *Honest Signals: How They Shape Our World* was published in 2008 by the MIT Press. His most recent book, *Social Physics*, was published by Penguin in 2014.