Article

Organizational Research Methods
1-23

DOI: 10.1177/1094428116677299
orm.sagepub.com

Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science

Scott Tonidandel[1], Eden B. King[2], and Jose M. Cortina[2]

Abstract

Advances in data science, such as data mining, data visualization, and machine learning, are extremely well-suited to address numerous questions in the organizational sciences given the explosion of available data. Despite these opportunities, few scholars in our field have discussed the specific ways in which the lens of our science should be brought to bear on the topic of big data and big data's reciprocal impact on our science. The purpose of this paper is to provide an overview of the big data phenomenon and its potential for impacting organizational science in both positive and negative ways. We identifying the biggest opportunities afforded by big data along with the biggest obstacles, and we discuss specifically how we think our methods will be most impacted by the data analytics movement. We also provide a list of resources to help interested readers incorporate big data methods into their existing research. Our hope is that we stimulate interest in big data, motivate future research using big data sources, and encourage the application of associated data science techniques more broadly in the organizational sciences.

Keywords

quantitative research, quantitative, multivariate analysis, philosophy of science, big data, data analytics, data science

As quantitatively minded scientist-practitioners, organizational scientists are ideally situated to help drive the questions and analytics behind the big data phenomenon. Advances in data science, such as data mining, data visualization, and machine learning, are extremely well suited to address numerous questions in the organizational sciences given the explosion of available data. For example, staffing practices, the ways that teams interact, and our understanding of leadership may be

[1]Davidson College
[2]George Mason University

Corresponding Author:
Scott Tonidandel, Davidson College, 209 Ridge Rd., Davidson, NC 28035, USA.
Email: sctonidandel@davidson.edu

revolutionized through the application of data science. More important, the big data phenomenon affords an opportunity to fundamentally change and improve our science.

Despite these opportunities, few scholars in our field have discussed the specific ways in which the lens of our science should be brought to bear on the topic of big data and big data's reciprocal impact on our science. To date, big data approaches have gained little traction in our literature or in the training of our students. Moreover, big data analyses can be criticized as a form of dust bowl empiricism (Landis, 2014; Ulrich, 2015). As a result, there is a substantial gap in our understanding of both the promise and perils of big data.

The purpose of this article is to provide an overview of the big data phenomenon and its potential for impacting organizational science in both positive and negative ways. In what follows, we define big data and discuss what we see as the key characteristics of the big data phenomenon. We then outline what we see as the biggest opportunities afforded by big data and discuss specifically how we think our methods will be most impacted by the data analytics movement. While the issues we discuss arise from advances in computer science and modern analytical methods applied to big data, we believe that many of the principles we are espousing can be applied to data sets of varying sizes. Finally, we identify some of the biggest obstacles to leveraging the power of big data alongside its potential pitfalls. Our hope is that we stimulate interest in big data, motivate future research using big data sources, and encourage the application of associated data science techniques more broadly in the organizational sciences.

What Are "Big Data"?

Defining big data can be tricky as there isn't one simple, agreed-on definition. Indeed, there is even disagreement over whether the term is singular or plural! Often, big data are defined according to the amount of certain attributes a data set has—often referred to as the Vs. The most basic of the Vs is *Volume*, which refers to the sheer size of the data set either in terms of the number of data points or disk space usage. When considering a data set's volume, sample size immediately jumps to mind, but high volume may arise from other sources such as a large number of measurements per individual. Another V of big data is *Velocity,* which refers to the throughput of the data (amount being added constantly) and the latency in using this info. A third attribute that describes big data is *Variety*, which refers to multiple data sources being integrated often with very distinct types of data being combined (e.g., numerical and text data).

Historically, these three Vs were seen as the defining features of big data (Laney, 2001). More recently others have added to the list of Vs. Sometimes the aspect of velocity that concerns the latency of data is given its own V—*Viscosity*. Another V that is commonly mentioned as a defining feature of big data is *Veracity*—the accuracy of the data. We believe accuracy is a critical element of all data, and thus we consider this to be the least compelling of the Vs when trying to define big data and differentiate it from a more traditional data set. Nevertheless, when the data contain substantial volume, variety, or velocity, it may be particularly challenging to ensure veracity. Thus, though we don't consider veracity to be a defining feature of big data, we recognize the veracity challenges inherent in big data.

While there is some apparent consensus that big data are characterized by the different Vs, there is disagreement regarding what constitutes a sufficient amount of each V (e.g., how many data points do you need for a data set to be considered large in terms of volume?) and whether the presence of all three Vs is necessary to constitute big data. Moreover, the targets for establishing baseline requirements for the three Vs are moving and might be discipline dependent. A data set that was large and would easily overwhelm available computing resources 20 years ago can be easily handled by the typical desktop computer or smartphone. Similarly, a computer science big data application may typically deal with hundreds of terabytes but a much smaller volume of data would easily overwhelm the available computing resources of most organizational scientists. Rather than focus on how much of each V might be present in a particular data set, we believe it is more productive to define data as "big

data" when characteristics of the data require you to change your mind-set. To the extent that one (or more) of the Vs requires you to think differently about your data, analyze them differently, embrace new technologies, statistical approaches, or theories, you have entered into the realm of "big data."

A Change in Mind-Set

While the Vs are an important lens through which we can view our data as big, the Vs are perhaps not as critical as the change in mind-set they require. In our view, the potential for big data to contribute to the organizational sciences lies not in the amount of any of the Vs nor in the combination of the Vs in one's data set. Rather, we see big data's revolutionary potential in terms of the new mind-set it can bring to research and its application to practice. Historically, much of our published literature has been conducted at a single point in time (or limited time frame) with a single well-defined group of participants, using a constrained set of variables measured in a certain way. In contrast, let's imagine a typical study that involves diverse types of data and multiple measures of phenomenon; the data are often being collected continually and updated with no specific start and end to the data collection; the analyses are being continually updated as new data come in. We believe that these properties of the typical data analytic study have important implications for how we conceive of our science and afford numerous opportunities for advancement.

The change in mind-set that we are espousing is perhaps best captured by Brieman (2001) where he describes two cultures of statistical modeling. The first culture, labeled *data modeling*, assumes that the response variable is generated from a given stochastic model. Data are collected and parameters are generated according to the assumed model, which is then validated using goodness-of-fit statistics. The contrasting culture is called the *algorithmic culture*, whose emphasis is on finding a function that describes the response process, and model validation is determined by maximizing predictive accuracy in cross-validation.

Brieman (2001) uses these two descriptors to criticize the current state of statistical science noting that the data modeling view, which he estimated that 98% of statisticians adhered to at the time of publication, was responsible for "irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems" (p. 199). We believe this perspective rings true if the word "statistician" is replaced by "organizational scientist." Our field's history, educational training, and journal policies all align with the data modeling culture as described Brieman. Perhaps unsurprisingly, our field has been criticized for theories that are never validated, findings that can't be replicated, and a divide between science and practice (Hambrick, 2007; Landis & Cortina, 2014). Putka and Oswald (2016) provide a compelling contrast of these two analytical cultures and how the organizational sciences could benefit by embracing the data algorithmic culture. They argue that the current data modeling culture inhibits our field's ability to more accurately predict valued outcomes, leads to models that fail to incorporate key drivers of a phenomenon, and can't adequately incorporate model complexity and uncertainty. Like Putka and Oswald, we believe that a more algorithmic culture may be helpful and big data sets and their corresponding analytical practices are one solution.

In summary, the term *big data* has been used interchangeably to refer to many different aspects of studies. Oftentimes, the term *big data* is used to refer to characteristics of the data source—the Vs. But, the term *big data* is also often used to describe a class of analytic approaches—big data methods. Finally, the term *big data* can describe an overall approach to problem solving that aligns with Brieman's algorithmic culture. These different interpretations of the term *big data* can lead to confusion. Unfortunately, it is often difficult to disentangle these three perspectives as there is considerable overlap among them. While big data sources can sometimes be analyzed with more traditional methods, the application of big data methods to those same data is often advantageous. But, those same methods can sometimes be profitably applied to data that would not be considered "big." Similarly, an algorithmic approach is perhaps more useful in conjunction with a big data source and big data methods, but that same mind-set may be beneficial in other situations as well.

Keeping in mind those various perspectives of big data, the following section addresses some of the potential opportunities and challenges we see for the organizational sciences.

Big Data Opportunities

An optimistic view of big data suggests that they generate exciting opportunities. While pessimists consider this excitement to be "hype" more than reality (Franks, 2012; Savitz, 2013), envisioning the possible avenues for organizational science is necessary to leverage its potential.

Opportunities to Investigate Old Questions in New Ways

One of the most interesting possibilities that emerges through the lens of big data is that organizational scholars can reexamine even the most straightforward conclusions of our science. Here we consider three examples of old questions that might be understood in new ways from this standpoint: (a) How can work be designed to maximize efficiency? (b) How can organizations select the best employees? and (c) How do people fit within their organization's culture?

A study of communication patterns in teams that used wearable personal sensors (i.e., sociometric badges; Pentland, 2012) to track network behavior linked the level of members' enthusiasm outside of formal work meetings to team-level performance. That is, the analysis of individual and team members' geospatial and verbal tracking data over numerous work days (with more than 100 data points per minute per person) pointed to the importance of social relationships in determining unit performance. This insight translated into recommendations to implement shared break times to facilitate (efficient) social connections, ultimately improving worker satisfaction and productivity.

The question of optimal selection (and classification) can be traced to the origins of IO psychology in the development of the Army Alpha and Beta tests in World War I and II. Yet, when we throw all of our best methods, tools, and tests (and corrections) at job applicants, the best we can do is explain 65% of the variance in job performance (F. L. Schmidt & Hunter, 1998), with most scholars finding that number to be extremely optimistic. It is reasonable to imagine that big data would allow us to explain some of the remaining variance of interest. This incremental validity may derive from better (or less collinear) predictors such as new forms of biodata (Putka, Beatty, & Reeder, in press) from Internet footprints (Youyou, Kosinski, & Stillwell, 2015), the development of realistic simulation assessments that monitor microexpressions and behavior, or genuine tracking of prior behavioral experiences. Improved validity could also come on the criterion side of the equation; perhaps our assessments of performance (an "old" question in and of itself) might also be improved by integrating large volumes of behavioral data from a variety of sources. Blending objective and subjective assessments of maximum and typical performance over time could allow for a much more comprehensive picture of an employee's genuine performance including the within-person variance that is typically found to outweigh the between-person variance that, traditionally, has been our focus (Ilies, Scott, & Judge, 2006). Linking these complex, multidimensional outcomes with dynamic predictors could allow for more complete models and, ultimately, increase validity (see Murphy & Shiarella, 1997). The magnitude of the incremental validity provided by big data and its utility given potential costs are important questions for organizational scientists to explore.

Person-organization fit and its impact on decisions to stay or leave an organization is a decades-old problem in the organizational sciences (Schneider, 1987). Cultural fit has been largely investigated as a static phenomenon, but data science techniques afford an opportunity for researchers to examine cultural fit as a dynamic constantly changing phenomenon. In a recent study, Goldberg, Srivastava, Manian, and Potts (2016) applied computational linguistics to over 10 million internal company emails to calculate acculturation trajectories for 600 employees over a 5-year period. They found that employees with slower enculturation rates had higher levels of voluntary turnover.

Table 1. Examining the Top Workplace Trends Through the Lens of Big Data.

Workplace Trend	Big Data Approaches	Potential Issues
Changes in laws may affect employment-related decisions	New models for adverse impact analyses	Unknown legal landscape
Growth of corporate social responsibility (CSR) programs	Sentiment analyses to determine internal and external stakeholders' immediate reactions to, and enable modification of, specific programs	Overly complex, limited ROI
Changing face of diversity initiatives	Leverage social media for targeted recruitment; assessing unconscious, subtle biases with passive data	Ethics; legal constraints
Emphasis on recruiting, selecting for, and retaining potential	Mining internal data to identify high potentials; continuous detection of initial withdrawal behaviors that may precede turnover	Privacy
Increased need to manage a multigenerational workforce	Individually adapted training environments	Need to integrate complex scientific knowledge in algorithm development
Organizations will continue to "do more with less"	Time use studies	Informed consent
Increasing implications of technology for how work is performed	Tracking technology use (time, location, efficiency)	May not integrate relevant contextual characteristics
Integration of work and nonwork life	Pairing wearable health and behavior monitors	Privacy
Continued use of HR analytics and big data	N/A—redundant!	
Mobile assessments	Immediate assessment and feedback systems	Reliability and validity of data; interpretations by end users

Importantly, these effects were found even after accounting for initial levels of person-organization fit. What mattered more for voluntary turnover rates was not initial fit but rather the rate at which individuals were able to adjust and adapt to an organizations culture. These examples represent a small, but expandable, set of ideas for understanding traditional questions in organizational research. There are also opportunities for addressing emerging practical concerns.

Opportunities to Address Emerging Practice Needs

In Table 1, we use the Society for IO Psychology's list of the top 10 workplace trends for 2015 as a vehicle for briefly describing where big data approaches may or may not be useful in addressing emerging practice needs (Below, 2014). Importantly, we also note potential issues and concerns that might arise.

In the case of the changing legal landscape, one area in which big data and modern analytics are relevant is in the evolution of adverse impact analyses. Jacobs, Murphy, and Silva (2013) noted that, to the extent that traditional tests of statistical significance replace traditional four-fifths calculations in adverse impact analyses, increases in sample size can correspond with inflated determinations of violations. Big data methods, such as those described later in this article, are needed to build new legal standards.

The need for and evaluation of corporate social responsibility (CSR) programs may also be studied using big data. Analyses of Twitter feeds in geographically relevant areas, for example, could help organizations design critical CSR initiatives in "real time." Yet, the complexity required

to acquire, analyze, and interpret such data may not be justified by the return on investment; similar information might be gained from newspaper headlines.

Efforts to improve diversity and inclusion can be informed by big data. Indeed, Botsford Morgan, Dunleavy, and DeVries (2015) argued that the subtle form of contemporary discrimination may only be adequately assessed through semi- or unstructured data such as email, sound and video clips, texts, and click streams. Such data—should privacy and ethical concerns be overcome—may be useful in identifying and disrupting problematic social networks within organizations. Algorithms could be developed that identify excluded employees and generate automatic invitations for formal (e.g., opportunities for development) or informal activities (e.g., happy hours).

Interest in "hi-pos" (i.e., high-potential employees) might also benefit from tailored algorithms that, for instance, map emerging leaders' behaviors against retrospective analyses of the small and big decisions (i.e., decision trees) of outstanding performers by mining internal data. In this way, the specific behavioral patterns that determine success in a particular context can be used to proactively identify employees with similar experiential trajectories.

Concerns that arise in light of the aging American workforce include age-related differences in training needs and learning styles (Beier & Ackerman, 2005). Big data techniques could be leveraged to build sophisticated adaptive training environments that cater learning experiences to learners' behaviors. Analogous to computer adaptive testing, big data-based programs might track the clicks and movements of learners and deliver maximally appropriate learning materials, assessments, feedback, and goals throughout the training experience. The effectiveness of such programs is inextricably tied to—and thus limited by—how well the driving algorithms are grounded in learning theory and evidence.

Time use studies, which traditionally rely on retrospective self-reports of time allocated to different tasks, could be automatized via computing technology. Understanding precisely how employees spend their time on a day-to-day basis could enable individuals or organizations to develop efficient structures and procedures. Such ideas have already been put into action: Big data techniques were used to determine that call center employees were more efficient when their breaks (i.e., social interaction time) overlapped (Waber, 2013). Moreover, these kinds of techniques can also be used to identify the technologies that support—rather than detract from—work.

The intersection between work and nonwork can also be studied through the lens of big data. For example, geospatial tools linked with biosensors might provide new insights on commuting patterns and stress. Personalized apps that help employees reach work and home in the most efficient, least stressful, and most productive manner could further serve as interventions to improve the work-nonwork interface (see http://time.com/3700921/apps-commute for examples).

Finally, the trend toward mobile assessments can be further enhanced through big data approaches. Continuous assessments could be paired, for example, with mobile continuous feedback platforms. Companies are already using feedback systems that nudge managers to provide immediate feedback (Streitfeld, 2015). Indeed, the *New York Times* reported,

> A new generation of workplace technology is allowing white-collar jobs to be tracked, tweaked and managed in ways that were difficult even a few years ago. Employers of all types—old-line manufacturers, nonprofits, universities, digital start-ups and retailers—are using an increasingly wide range of tools to monitor workers' efforts, help them focus, cheer them on and just make sure they show up on time. (Streitfeld, 2015)

Opportunities to Investigate New Questions

Big data also allow for entirely new questions. Criticized as "dust bowl empiricism," big data nonetheless offer virtually limitless opportunities to observe genuine patterns of relations. Kitchin

(2014) contrasts a theory-free empiricist view of big data with the less extreme view that big data offer opportunities for applying the tenets of the scientific method to combinations of inductive and deductive approaches. In the latter, scientists use "guided knowledge discovery techniques to identify potential questions (hypotheses) worthy of further examination and testing" (p. 6). In the social sciences, in particular, big data can blend wide-scale and finer-grained analytic approaches by providing information about individual behavior within and across contexts.

Indeed, one area in which big data may have a particularly large impact is in building understanding of how the social systems in which organizations are embedded influence work and workers. The interest of organizational scholars in cross-cultural psychology has grown with the globalization of work, yet much is unknown about the ways that elements of culture, time and place shape organizational behavior. This lack of understanding is, at least to some extent, attributable to the challenges of multinational data collection. Big data offer opportunities to develop new ideas about these dynamics.

For example, Chung and Pennebaker (2012) summarized an emerging body of evidence that applies natural language processing methods to questions about culture and community. In their view, written language is a "window" into culture and emerging events in which linguistic or sentiment analyses of social media platforms such as Twitter or Facebook serve as representations of behavior, belief, or feelings. In one domestic study, Hernandez, Newman, and Jeon (2016) analyzed the Twitter feeds from the 200 largest cities in the United States. By focusing on phrases related to "loving" or "hating" a job, Hernandez and colleagues found that the content of job-related tweets from people in cities with high SES, occupational prestige, and commute times was more negative than other cities. The consideration of "macro-level" attitudes and behavior that is enabled by big data opens up entirely new research questions and the potential for novel streams of work.

More broadly, popular organizational theories can help to drive new big data investigations. For example, popular theories of performance (e.g., Campbell, Dunnette, Lawler, & Weick, 1970), team dynamics (e.g., Marks, Mathieu, & Zaccaro, 2001), and stress (e.g., Karasek, 1979) can give big data analysts jumping off points for prioritizing, organizing, and interpreting findings. As a specific example, Waber (2014) drew from influential theories of sex and gender to study the behaviors of men and women across industries and organizations in the United States using massive amounts of raw data in three different companies over six-week time frames. He considered the broad (and socially critical) question of whether there are robust differences in objectively quantified workplace behavioral patterns of men and women that might account for gender differences in income and promotions. The detection of sex differences in behavior has been studied for a century, and opposing conclusions are drawn from meta-analyses that imply either that women's behaviors drive their status or that no gender differences exist (see Barnett & Rivers, 2009; Eagly & Wood, 2013). What is unique about Waber's approach to testing the competing hypotheses is that he used objective indicators of actual behavior including emails, phone logs, instant messages, speed of transactions (an indicator of productivity), and speech patterns. The only gender difference that emerged in any of these actual behaviors was the speed of transactions—women who worked in call centers processed calls more efficiently than men. This study exemplifies the application of big data tools to test dominant theories in new ways that ultimately improve our science.

Opportunity to Fundamentally Improve Our Science

Despite its many strengths, our science can be criticized on a number of important fronts. The social sciences currently face what has been termed a replication crisis. Very few studies attempt to replicate past findings. Figure 3 from Colquitt and Zapata-Phelan (2007) illustrates a disturbing trend in our literature where the amount of actual theory testing is dwarfed by the number of studies building new theories or expanding on existing ones. This paucity in theory testing can be partly attributed to a preference displayed by virtually all journals for novelty over replication. In the rare circumstances when replications are attempted, numerous research findings across many different

disciplines fail to replicate. Several fundamental problems with traditional research designs may be contributing to this replication crisis. Historically, most studies are executed on a relatively modest number of participants using a limited set of participants that may not generalize to other settings. A continual complaint has been proffered for decades that the typical study is underpowered (e.g., Cohen, 1992; Maxwell, 2004). While the most obvious consequence of this is the occurrence of Type II errors, an often overlooked corollary of this phenomenon is that the magnitude of published effects is often overstated, because the observed effect must be biased upward to be statistically significant leading to publication bias. Similarly, the limitations of statistical significance testing have been discussed extensively in many sources. Despite the widespread acknowledgement of these problems, very little change is appearing in terms of our reliance on these techniques.

The era of big data affords us the opportunity to counter these criticisms with better science. The data science perspective is founded on a model of replication with the potential for theories and models to be tested and retested, verified, and updated as additional data, either in terms of volume, variety, or velocity, is available. In fact, replication is a requirement of sound data science. Results stemming from data analytic techniques that have not been validated on additional samples, would likely possess very little, if any, traction—and deservedly so. Importantly, these replication and cross validation efforts are not one-shot endeavors but rather continue over time, providing ongoing and more nuanced information regarding the continued utility of a past finding. In addition, concerns about power and statistical significance become largely moot when the sample size approaches that of most big data studies. For example, a study with a sample size of $N = 3{,}000$ will have greater than 80% power to detect the effect of a variable that explains $.025\%$ of the variance in another variable. This 80% power for explaining as little as one tenth of 1% of the variance in a variable can be achieved with a sample smaller than $N = 8{,}000$. Both of these sample sizes are not considered large in a big data sense, but this nonetheless illustrates the fact that even very small effect sizes will be found to be statistically significant in most big data studies. As a result, big data analytics affords the potential for shifting our collective behavior toward alternative techniques, such as a focus on precise parameter estimation (Maxwell, Kelley, & Rausch, 2008), because traditional significance testing is often uninformative in very large data sets.

How Big Data Analytics Will Impact Our Methods

Learning to Live Without a "True" Model

Multiple regression is arguably the most widely relied on statistical approach in the organizational science. Multiple regression and its many generalized linear model derivatives (logistic regression, structural equation modeling, multilevel modeling) all assume that the correctly specified model is being used for the analysis. Unfortunately, our theories are seldom sufficiently developed to include all or even most relevant variables. Moreover, we often don't even know what those missing variables may be. Thus, we often find ourselves testing a very limited set of variables, and our failure to incorporate critical omitted variables can have important implications for the accuracy of the conclusions we may wish to draw from our data (Antonakis, Bendahan, Jacquart, & Lalive, 2010).

When we do have many variables, we first use them to try to identify the one best model and then interpret the results from that single chosen model rather than considering how a more robust set of variables might be relevant. Importantly, this approach fails to acknowledge the uncertainty that exists regarding our choice of the "best model." This scenario is perhaps best described by Hoeting, Madigan, Raftery, and Volinsky (1999):

> Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data.

> This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are. (p. 382)

In contrast to this typical approach, the data analytic approach relies on multiple models or model ensembles. Rather than select the single best model and assume that it correctly describes the data generation process, model ensembles analyze all of the possible models that can be derived from the existing set of variables and aggregate the results using a variety of techniques (e.g., bootstrap aggregation/bagging, boosting, Bayesian model averaging, etc.— for a review of ensemble methods, see Rokach, 2010; Seni & Elder, 2010). The resulting ensemble models consistently outperform the selected "best" model producing superior prediction accuracy across a variety of domains and ensemble techniques (Burnham & Anderson, 2004; Kaplan & Chen, 2014; Markon & Chmielewski, 2013; Opitz & Maclin, 1999). It should be noted that more advanced data analytic techniques will not always outperform more familiar techniques (for examples, see Collins & Clark, 1993; Schmidt-Atzert, Krumm, & Lubbe, 2011).

An underappreciated advantage of incorporating uncertainty into our models is that they can actually help improve on existing theory by identifying new meaningful variables we may not have otherwise considered. By learning how the multiple models combine to create the ensemble models and boost prediction, there is the potential that we could identify new variables that should be part of our theories but currently aren't. Big data can be a source of discovery. It can provide "Eureka" moments and insights into relationships and patterns we would have never known to look for (Dyche, 2012). Hsieh, Hung, and Ho (2009) provide a demonstration of how an ensemble classifier model can be used to identify association rules between applicants' characteristics and their credit worthiness, enabling them to develop new strategies for credit approval analysis. The authors first demonstrated that an ensemble classifier out performed multiple individual classifier models in the prediction of credit worthiness. They then used a Markov blanket to examine both how the predictive features relate to the target variable of credit worthiness and also how the features themselves relate to one another. Using this knowledge, the authors were able to condense the entire ensemble down to seven of the most informative nonoverlapping association rules.

Moving Beyond Linearity

Modern data analytic modeling methods will also benefit our science because these methods often eschew traditional assumptions of linearity and instead allow us to more easily incorporate nonlinear relationships. Perhaps more important, they are better suited for capturing discontinuous change. Steve Guastello and others (e.g., Guastello, Koopmans, & Pincus, 2009) have been developing catastrophe modeling techniques for decades, but catastrophe modeling hasn't really caught on. Many modern data analytic methods such as random forests, artificial neural networks, and support vector machines are ideally suited for identifying both linear and nonlinear patterns in the data. Importantly, the detection of nonlinearity and interactions using these techniques can happen without prespecification of the exact pattern of nonlinearity. Numerous studies have found that artificial neural networks often outperform traditional regression based techniques in terms of their ability to predict counterproductive work behaviors (Collins & Clark, 1993), employee health (Karanika-Murray & Cox, 2010), turnover (Somers, 1999), and job satisfaction and performance (Somers & Casal, 2009). In all of these instances, the superiority of the artificial neural network was attributed to its ability to accommodate nonlinearity, even when compared to regression approaches with nonlinear terms included. Of course, the identification of these new nonlinear patterns should be viewed as suspect until they have been cross-validated. In each of the examples above, the models incorporating nonlinearity were all validated on a holdout sample that was not used to train the neural network. Big data, because of volume and velocity if nothing else, may make the existence of discontinuous change more transparent and, therefore, more easily studied.

Renewed Emphasis on Measurement

For decades, organizational researchers have lamented the limited nature of the criterion being used in most analyses (e.g., the criterion problem). While the original sentiment concerned the limited ways in which job performance is defined and measured, we believe the overarching critique of the limited conceptualizations of our variables is applicable to a variety of domains. The big data analytics movement affords an opportunity to expand beyond these narrow conceptualizations by: aggregating data from multiple sources, incorporating longitudinal/dynamic measurement, and using biological indicators of constructs.

Because we in the organizational sciences so often deal with psychological constructs, we receive more training in measurement than do most scientists. Indeed, the early days of our field contained more work on measure development and evaluation than on any other topic. The earliest issues of the *Journal of Applied Psychology* were rife with articles on cognitive ability measurement (e.g., Jarrett, 1918; Pintner & Toops, 1917; Yerkes, 1917). Soon thereafter, we branched out into interest measurement (e.g., Freyer, 1930; Strong, 1951, 1952), and personality (McKinley & Hathaway, 1942; H. O. Schmidt, 1945), to name a few.

Technological (and theoretical) advances have brought new measurement challenges. For example, the ubiquity of smartphones has made experience sampling much more feasible. At the same time, respondents cannot be asked to complete long surveys several times per day. Well, they can be asked, but they will demur. Thus, we have had to find the right balance between length and content on the one hand and attentional capacity of respondents on the other. And it is vital that we do so because experience sampling has opened up new realms of possibility (e.g., the majority of variance in contextual performance is within person; Ilies et al., 2006).

The aforementioned study by Goldberg et al. (2016) illustrates some of the measurement possibility with big data. That study leverages data variety by using a new kind of data (email streams) to assess onboarding and velocity by incorporating the dynamic nature of that continuous process into their hypothesis tests. Other examples of new kinds of measurement resulting from big data can be found in the personality domain. Concerns around the accuracy of personality assessment due to faking and other potential response distortions have existed for quite some time. Improvements in the assessment of personality may be derived by integrating new and different sources of measurement. Park et al. (2014) describe a language processing model using social media data from almost 70,000 Facebook users. Their results point to this being a valid approach for assessing personality that can provide incremental validity over traditional techniques. Youyou et al. (2015) also explore new measures of personality using social media data, but instead of relying on posts, they use Facebook "likes" (i.e., marks that represent a Facebook user's evaluation of others' posts). They found that in some instances the computer-derived personality judgments had higher criterion validities than self-reported personality.

What measurement challenges will be created by big data? It is impossible to know for sure, but the combination of variety and velocity if nothing else will create new problems. How do we extract meaning from information coming from multiple, disparate sources? In particular, how do we do this on a second by second basis? How does this impact our traditional measurement concepts? With high-velocity big data, for example, test-retest reliability may no longer be a meaningful concept. For example, if one possesses moment-by-moment measurements over a long span of time, the average of any random error in those measurements would essentially trend toward zero across such a large number of repeated measurements making test-retest reliability less relevant. Who better to tackle these problems than those of us with extensive training in measure development, scoring, norming, reliability, validation, etc.?

Iterating Inductive and Deductive Processes

The rise of big data has been accompanied by a new epistemological approach that seeks to develop knowledge from that data as opposed to using the data to test existing theory. This has laid the

foundation for headline grabbing claims that "theory is dead" (Anderson, 2008). Data analytics and their emphasis on discovery are often criticized as dust bowl empiricism. While the application of these methods can be atheoretical, it doesn't have to be. In fact, we would argue that an important contribution of the big data movement is that its methods will actually help us build better theories.

Current quantitative methods fail to capture both the complexity and the subtleties of phenomena of interest. In response, qualitative methods have assumed a more influential role in developing theory using an inductive process. Qualitative methods are valued for their ability to accumulate more in-depth information about a phenomenon, examine patterns over time, consider contextual information, and incorporate textual and other non-numeric sources of data. In contrast, quantitative methods are recognized for their ability to apply statistical analyses to identify relationships between variables, provide measurable evidence for phenomenon, and test theories.

Given the different strengths inherent in these two approaches, researchers have tried to capitalize on both by integrating these two procedures into mixed methods research. The benefits of using mixed methods to develop richer and more complete theories and to evaluate the veracity of those theories seem obvious. Perhaps less obvious is that big data analytics can be characterized by many of the features of both the qualitative and quantitative approaches listed above. Big data accumulates more in-depth information, examines patterns over time, and incorporates textual and other non-numeric sources of data, while simultaneously applying statistical analyses to identify relationships between variables, providing measurable evidence for phenomenon, and testing theories. Thus, while we improve our predictive capabilities, we simultaneously improve our theory by incorporating new variables, testing out alternative models, replicating our results, and permitting our models to modify and change over time as needed. The application of modern modeling techniques to big data essentially transforms the validation process to one of continual validation as opposed to more static cross validation.

Big data is not free from theory nor should it be. Instead of equating big data to dust bowl empiricism, we believe that big data can create what Kitchin (2014) calls data-driven science. The model of data-driven science is one where data is used inductively in an iterative process to inform hypothesis generation and theory creation. While dust bowl empiricism comes to a halt after the data have been mined, the process of data-driven science continues by pairing the inductive method with deductive approaches. One can envision a new class of big data experiments that drive the creation of new knowledge (e.g., Bakshy, Eckles, & Bernstein, 2014). For example, a large European retailer text mined over 36,000 open-ended employee suggestions. The results of this analysis are now being used to develop new employee wellness initiatives that are being tested experimentally at the store level. Similarly, Google's HR analytics function is well known for using analytics to gain employee insights that are subsequently tested experimentally (Derose, 2013).

Importantly, the inductive process need not start in a theory-less vacuum. Existing knowledge can be used to shape the analytics engine to direct the process of knowledge discovery in order to derive useful conclusions rather than identifying any-and-all potential relationships (Kitchin, 2014). For example, the dictionary that underlies a text mining algorithm can be built around existing theory. Then the machine learning algorithm can build on the existing theoretically driven dictionary to identify previously unidentified terms. Indeed, the algorithm can ultimately improve on the existing theory by identifying word combinations or phrases that make use of terms from the original dictionary but improve one's predictive capabilities over knowledge of the single words alone. Similarly, when using random forests, it is possible to start with theory and build in top-down constraints into the tree building process. Thus, prior theoretical knowledge or situational considerations can be (a) incorporated from the start and (b) expanded through machine learning.

These basic concepts have been extended recently by Yan and colleagues (Yan, McCracken, & Crowston, 2014a, 2014b). These authors describe new processes that combine machine learning algorithmic coding of textual data with simultaneous incorporation of feedback from human coders. Taking it a step further, Landers, Brusso, Cavanaugh, and Collmus (2016) describe a process of

theory-driven web scrapping that follows a hypothetico-deductive approach. They propose designing a web-scrapping approach only after the theory and hypotheses have been determined using a data source theory to guide data collection.

As another example of how theory can be integrated into these methods, consider artificial neural networks. While it is true that these techniques are often applied with the sole purpose of improved prediction, this doesn't have to be the case. Despite being one the most "black-box" analytic techniques, neural networks can be leveraged to improve and enhance current theory. For example, Somers (1999) used neural networks to study turnover and found pronounced floor and ceiling effects (nonlinearity) for a number of variables including job satisfaction, withdrawal intentions, and affective commitment. The implication of this finding for theory is quite profound in that it suggests that changes in these potential antecedent variables will have little to no relationship to turnover across large ranges of the antecedent variables. But at certain thresholds values, small changes can be quite important for predicting future turnover. This suggests that in order to understand turnover better, we must try to understand the catalyst events that push individuals across the thresholds. Other examples of how neural networks can contribute to theory can be found in Somers and Casal (2009) and Somers (2001).

One of the mistaken assumptions about big data is that it somehow replaces existing techniques (whether qualitative or quantitative) or that these different methods are mutually exclusive. A better perspective might be to view all of these approaches as complementary (Blok & Pederson, 2014). Each alone fails to provide a complete solution to a problem. But, they can be integrated in an iterative process to improve our scientific techniques.

Though one of the common criticisms of data analytics is the data-driven nature of the approach, deriving knowledge from data is not inherently bad. As we have already discussed, using big data inductively has the opportunity to help improve our theories rather than abandon theorizing altogether. The key is to recognize the limitations inherent in such an approach and to consider pairing such an approach with more traditional deductive methods. Moreover, we believe that big data offers a level of transparency that doesn't currently exist with regard to post hoc theorizing. Current analytical practices for non–big data studies are similarly data driven in many instances, but this goes unacknowledged. Numerous published findings and their corresponding theories are likely derived from the data via HARKing—hypothesizing after the results are known (Kerr, 1998). Kerr and Harris (as cited in Kerr, 1998), in a survey of 156 behavioral scientists, revealed that two HARKing behaviors occurred just as frequently as the hypothetico-deductive approach. This suggests that data-driven analyses and inductive theorizing is occurring with reasonable frequency in the traditional quantitative study. Criticizing data analytics for engaging in the same inductive behaviors currently being practiced in traditional quantitative analyses seems unjust. Instead, data analytics should be applauded because the data-driven aspects of the process are explicit, as opposed to never acknowledged like most HARKing behaviors. Moreover, the data analytic approach is predicated on cross-validation and replication—again a feature missing from some traditional studies.

New Perspectives on Sampling

The era of big data raises a number of interesting and important questions about sampling. On the one hand, an argument might be made that sampling and statistical theory based on sampling is irrelevant as one can simply analyze the entire population. With big data technologies, the marginal cost of additional data collection approaches zero making the need for sampling less relevant. However, we would argue that importance of sampling endures and that big data affords an opportunity to once again think critically about sampling and look at more nuanced conceptualizations.

Despite having large samples, the question remains whether the big data "population" is the correct population for the question being investigated. While big data is often described as having a sample size equal to the population (e.g., $N =$ all), the available data may not be representative.

Importantly, a large *N* does not mean a representative sample. One potential positive outcome of big data is that it reemphasizes sampling issues. Most conventional studies are currently carried out on convenience samples. Yet, we neglect this fact when we apply parametric procures such as null hypothesis significance testing and confidence intervals. Estimates of sampling error should be viewed with caution and require replication under such conditions. One marked improvement resulting from big data is that replication and cross validation of findings are inherent requirements of data analytic studies. But an understanding of the representativeness of any big data sample is of critical importance. Our hope is that big data permits us to move away from the mind-set of the traditional study where we are just happy to have data from a sample to one where we ask more critical questions about the appropriateness of the data at hand.

Another aspect of sampling that is neglected in most modern studies is time. The goal of many studies is to make predictions into the future but samples are often collected at a single point in time. Big data affords an opportunity to bring the time dimension of sampling to the forefront. The velocity of the data permits the continual testing and validation of results on new samples obtained on a rolling basis. Moreover, big data and data streams permit the application of numerous sampling methodologies that the typical organizational scientist is unfamiliar with such as stream sampling, hash sampling, and graph sampling to name a few.

Finally, big data allows one to investigate new samples that previously couldn't be analyzed. With extremely large data sets, one can now investigate minority populations in a way never before possible. For example, Welles (2014) describes her work on female online gamers. In one instance, she was interested in examining a very specific group of high frequency female gamers over the age of 50. In a database of 10 million Second Life users, only 1,500 met selection criteria for the population of interest. Thus, she was able to leverage an extremely large data set to make a detailed examination of the gaming behaviors of an extreme statistical minority with a meaningful sample size. The flip side of this example is that any holistic analysis of the entire data set would totally overwhelm the information being provided by this critical minority group. Because the massive size of the original data results in meaningful sample sizes for statistical minorities, big data affords an opportunity to focus more critically on segmentation. Smaller subsets of observations can be analyzed individually to hopefully generate valuable insights for those specific groups.

Challenges for the Organizational Sciences

Data Integrity

The inclusion of more multifaceted measures can create issues surrounding data integrity and a concern commonly expressed as the Garbage-In, Garbage-Out phenomenon. However, given our field's expertise in measurement, we are well positioned to contribute to the big data movement in this critical area. Consider the numerous technological advances that have been made in the wearable sensor space such as sociometric sensors. These devices exploit the fact that many people are already comfortable with wearable electronics (such as cell phones, digital watches, pedometers). A variety of highly accurate data can be available such as nonlinguistic social signals (such as interest, excitement, influence) and relative location monitoring. Such sensors are being used to investigate a host of phenomena in organizational behavior. For example, activity and number of team interactions has been shown to be related to creativity (Tripathi & Burleson, 2012). Similarly, Olguín and Pentland (2010) found that activity level and interaction patterns as measured by sociometric sensors predicted success by teams in an entrepreneurship competition.

These devices clearly have a number of benefits over traditional observational data collection methods and can replace costly human observation, susceptible to subjective biases and memory errors. But the measurement properties of these devices are largely unknown. Chaffin et al. (2015)

describe some of the challenges with regard to using wearable sensors in organizational studies. Take the case of a simple network analysis using sociometric sensors. These sensors use Bluetooth technology to record when team members are in each other's presence. However, it is unlikely that all badges are manufactured to such a specification that they are equally sensitive. As a result, the individual with a more sensitive sensor will likely record more team member's signals. As another example, a sensor might identify two people on either side of a wall as being in each other's presence, while two people at either end of a conference table wouldn't be. With repetition, these measurements will create the appearance of being more central in the network.

An additional concern with these the types of measures used in big data studies that is often ignored is construct validity (Lazer, Kennedy, King, & Vespignani, 2014). Though interesting relationships can be found, clear evidence regarding construct validity is sometimes lacking. For example, though Goldberg et al. (2016) found evidence that the linguistic similarity between an individual's incoming versus outgoing email communications within the organization was related to less voluntary turnover, it remains to be seen whether these email patterns are really capturing enculturation and fit.

The key issue here is that we don't know whether this is happening with this technology or any other big data collection method, but we possess the requisite skills and models to investigate these issues. We can play a critical role in enhancing the veracity of the information produced by these and other devices by applying our theories of measurement to the testing and evaluation of these products.

Inadequate Preparation for Big Data Projects

Sharpe (2013) cites inadequate statistical education as a major culprit in inhibiting the adoption of better data analytic practices. Aiken, West, and Millsap (2008) support these assertions from a large scale survey of PhD psychology programs where they observed a lack of statistics training. More recently, Tett, Walser, Brown, Simonet, and Tonidandel (2013) identified similar deficiencies in PhD programs in I-O psychology and management. To further complicate the issue, big data will require training in certain areas in which we now receive no training, and other areas in which we receive less training that we used to. Most of us receive no training in qualitative inquiry. We therefore lack the knowledge and skills necessary to conduct mixed methods research. Moreover, data analytics generally relies on software with which we are entirely unfamiliar (e.g., Hadoop, MapReduce, Python) and statistical techniques we have no exposure to (e.g., machine learning, support vectors, neural networks). As was mentioned earlier, big data also creates more and novel measurement challenges at a time when training in measurement is being replaced by more specialized substantive and data analysis training.

These realities may force us to do more of our work with multidisciplinary teams. Unfortunately, our field has done a poor job in terms of cross-fertilization with our other major research domains (Allen, 2015). Expanding our capabilities on this front will be key to making an impact in the big data era. Some disciplines have more familiarity with qualitative methods than we do, some have more exposure to data analytic techniques for big data, and some have more experience in managing diverse databases and integrating information across platforms. We can bring to the table scale development and validation and quantitative analysis techniques. If nothing else, however, we will have to learn how to speak the language of mixed methods and of data analytics.

Interpretability and the "Black Box"

Admittedly, some of the data analytic techniques applied to big data make interpretation of the underlying meaning of particular models quite difficult. For example, artificial neural networks are used to create a series of layers and nodes that can be used to make predictions from a set of variables. One common criticism of these techniques is the apparent ''black box'' nature of the networks that are created: ''Neural networks are best approached as black boxes with mysterious inner workings, as

mysterious as the origins of our own consciousness" (Berry & Linoff, 1997, p. 282). While neural networks can provide impressive improvements in the accuracy of predictions, the meaning of the different nodes and layers and the nature of the relationship between the predictors and criterion are often difficult to surmise. While this is a valid criticism, tools for interpreting neural networks (Intrator & Intrator, 2001), including visualizing the neural network fitted functions (Plate, Bert, & Band, 2000), are being developed to increase the utility of these approaches. Somers (1999, 2001; Somers & Casal, 2009) provides some nice examples relevant to the organizational sciences illustrating how neural networks can be used to gain insight for improving theory.

One can also disparage big data studies for the over interpretation of a particular finding. However, we would point out that similar condemnations can be levied against any statistical approach or methodology. Concerns about erroneous conclusions of causality, overgeneralization of results, and the misapplication of a research finding is not the fault of the methodology. Such criticisms are not limited to data analytic studies but are relevant to all studies and thus are not emblematic of a central problem with the big data analytic techniques themselves.

Another concerning aspect of the black-box nature of many algorithmic applications is the lack of algorithmic transparency. In some cases, those developing and using algorithms fail to disclose the exact nature of the algorithm citing proprietary concerns. In other instances, it may even be unclear whether the exact nature of the algorithm is even known. For example, O'Neil (2015) describes the value-added teacher model that uses an algorithm to measure teacher effectiveness and the inability of various constituencies to actually articulate the nature of the algorithm in any meaningful way. Chaffin et al. (2015) report similar challenges with sociometric sensors that output numerous metrics, but the hardware manufacturer refuses to share the specifications of the algorithm that produces the output. As a result, one is left to blindly accept the veracity of the results that are provided by the algorithm.

Ethical Issues

Concerns about data privacy have been discussed extensively in the media and other various outlets. Rather than reiterate those points, we would instead like to focus on some ethical issues specifically related to big data research that receive less attention. Big data often are not collected with a specific purpose in mind but are instead just a by-product of the continual collecting and storing of existing information. An additional complication arises from our ability to link multiple stores of information together. This poses a number of ethical challenges. Informed consent can't be obtained because the purpose of study, or even the intention to study at all, is perhaps unknown prior to the start. Anonymity cannot be guaranteed because identity can be determined from so many different pieces of information and because linking of observations across data sources requires identification. Unique identifiers are not a solution to this problem as surprisingly few pieces of information are actually required to identify specific individuals. For example, Sweeney (2000) demonstrated how anonymized health records could be joined with publicly available municipal voter roll data to identify individuals. She found that 87% of the U.S. population can be uniquely identified from just three pieces of seemingly anonymous data: zip code, gender, and date of birth. It is also difficult to ensure minimal risk because data that are minimal risk in and of themselves could be linked to other data sources; the exposure of which might represent greater than minimal risk, as illustrated by Sweeny and her ability to identify anonymous health data. Moreover, machine learning algorithms may detect a pattern that could put a person at risk that was not conceived of prior to the implementation of the algorithm. It may be impossible to fit these issues into the existing IRB framework.

Another ethical risk comes from what is sometimes termed machine bias (Angwin, Larson, Mattu, & Kirchner, 2016). While data are often thought of as objective, big data and the algorithms arising from such data may not be. In a recent white paper released by the White House, Muñoz, Smith, and Patil (2016) present salient examples of how poorly selected, incomplete, incorrect, or

outdated data can all result in discriminatory outcomes. For example, an algorithm developed on existing data to maximize employee fit could very easily just be perpetuating historical biases that exist in the data set. Similar illustrations have been offered in other areas such as the risk-assessment algorithms that are used for criminal sentencing (Angwin et al., 2016). As a result, extreme care must be taken to strive for "equal opportunity by design" (Muñoz et al., 2016, p. 6). This principle argues that from development through all stages of an algorithm's life, fairness and equal opportunity must be built into a system's design and safeguards need to be in place to protect from discrimination.

Conclusion

The era of big data has arrived. The question is how we should respond. Big data are not a magical panacea. There are shortcomings to be acknowledged and caution to be exercised. But there is also a tremendous amount of opportunity to be realized. This opportunity comes not only in the form of new research questions, but also in terms of our methodological approaches. In an effort to aid those individuals who wish to leverage big data capabilities, we have compiled a set of resources to assist in the implementation of these approaches (see Tables 2-5). These accompanying tables are by no means a comprehensive list, but should nevertheless help interested readers incorporate big data methods into their research. It is our hope that our field embraces the benefits afforded by these data analytic approaches to improve our science.

Table 2. R Packages for Big Data.

Package	Description
Social media	
SocialMediaMineR	Number of hits of URLs on social media
twitteR	Interface for Twitter API
Rfacebook	Interface for Facebook API
tumblR	Interface for Tumblr API
Rlinkedin	Interface for LinkedIn API
Rflickr	Interface for Flickr API
Text mining	
Tm	Comprehensive framework for text mining applications
zipfR	Word frequency distribution analysis
RTextTools	Machine learning package for automatic text classification geared toward social scientists
textir	Multinominal logistic regression for phrase counts
lsa	Latent semantic analysis
NLP	Basic methods for natural language processing
Visualization	
ggplot2	Complex plots using grammar of graphics
bigvis	Data plots using aggregation and smoothing techniques
tabplot	Visualizations of multivariate data sets
googleVis	Interface for Google Charts API
threejs	3D scatterplots and globes
wordcloud	Horizontal and vertical visualization of text with font size based on word frequency
Parallel computing	
snowfall	Simple parallel computing
pdbMPI	Single program/multiple data parallel computing
parallel	Coarse-grained parallelism
Rmpi	Interface to MPI

(continued)

Table 2. (continued)

Package	Description
Machine learning and data mining	
FactoMineR	Multivariate exploratory data analysis
cluster	Basic clustering techniques
rattle	Graphical user interface for data mining algorithms
arules	Market basket analysis and association rules
nnet	Neural networks
randomForest	Classification and regression using random forest algorithms
bigrf	Random forest computations in parallel for data sets too large for storage in memory
gbm	Gradient boosting
rpart	Recursive partitioning and regression trees
e1071	Functions for support vector machines, bagged clustering, and naïve Bayes clustering
glmnet	Lasso and elastic net regularized generalized linear models

Table 3. Python Packages for Big Data.

Package	Description
Social media	
python-twitter	Interface for Twitter API
facebook-sdk	Interface for Facebook API
pytumblr	Interface for Tumblr API
python-linkedin	Interface for LinkedIn API
Flickrapi	Interface for Flickr API
Text mining	
textmining	Statistical text analysis
NLTK	Natural language processing
gensim	Latent semantic analysis
Visualization	
Matplotlib	2D plotting library
Bokeh	Visualizations designed for optimal web browser viewing
Mayavi	3D visualization
pygooglechart	Interface for Google Charts API
plotly	Plotting library for graphs
Parallel computing	
dispy	Single instruction/multiple data parallel computing
papyros	Master-slave based parallel processing
Jug	Task-based parallelization
mpi4py	Interface to MPI
Machine Learning and Data Mining	
scikit	Set of modules for machine learning and data mining
Orange	Component-based data mining
mlpy	Machine learning library
pybrain	Neural networks
milk	Machine learning toolkit focused on supervised classification

Table 4. Other Resources for Big Data.

Resource	Description
Amazon Machine Image	Cloud computing through Amazon Web Services
Deeplearning4j	Open source distributed deep-learning library written for Java and Scala
Funf Open Sensing Framework	Open source tools for creating sensing applications for mobile phones
Hadoop	Distributed storage and processing of large data sets across clusters of commodity servers
IBM Watson Developer Cloud	Suite of tools including text tone and emotion analyzers, natural language classification, image analysis, and more
Kdnuggets.com	Contains resources, news, software, and tutorials related to business analytics, big data, data mining, and data science
Microsoft Azure	Cloud computing through Microsoft's cloud platform
Rdatamining.com/big-data	Tutorials and resources for big data platforms
Revolution Analytics	Version of R for big data statistical analysis
Tableau	Data visualization software focused on business intelligence
Waikato Environment for Knowledge Analysis (Weka)	Free suite of visualization tools and machine learning algorithms for data analysis and predictive modeling

Table 5. Tutorials on Using R and Python for Big Data.

Description	Tutorial
Analyzing Social Media Data in R	http://thinktostart.com/category/datascience/r-tutorials/
Big Data Resources for Python and R	https://www.datacamp.com/community/tutorials/learn-data-science-resources-for-python-r#gs.TYnrlYM
Classification Trees in R and Python	https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/#nine
Cluster analysis in R	http://www.stat.berkeley.edu/~s133/Cluster2a.html
FactoMineR Tutorial	https://www.youtube.com/playlist?list=PLnZgp6epRBbTsZEFXi_p6W48HhNyqwxlu&feature=view_all
ggplot2 in R Tutorial	http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html
Machine Learning in R	https://www.datacamp.com/community/tutorials/machine-learning-in-r#gs.kTLLgDA
Natural Language Processing using NLTK in Python	http://www.nltk.org/book/ch01.html
Neural Networks in R	http://www.di.fc.ul.pt/~jpn/r/neuralnets/neuralnets.html
O'Reilly Course on Data Visualization using Python	http://shop.oreilly.com/product/0636920046592.do?sortby=bestSellers
Parallel Computing in R using snowfall	http://www.informatik.uni-ulm.de/ni/staff/HKestler/Reisensburg2009/PDF/snowfall-tutorial.pdf
Support Vector Regression in R	http://www.svm-tutorial.com/2014/10/support-vector-regression-r/
Text Mining in R	https://rstudio-pubs-static.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*, 32-50. doi:10.1037/0003-066X.63.1.32

Allen, T. D. (2015). Connections past and present: Bringing our scientific influence into focus. *Industrial-Organizational Psychologist*, *52*(3), 126-133.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*. Retrieved from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *Leadership Quarterly*, *21*, 1086-1120. doi:10.1016/j.leaqua.2010.10.010

Bakshy, E. B., Eckles, D., & Bernstein, M. (2014, April). *Designing and deploying online field experiments*. Paper presented at the 23rd World Wide Web International Conference, New York, NY.

Barnett, R., & Rivers, C. (2009). *Same difference: How gender myths are hurting our relationships, our children, and our jobs*. New York: Basic Books.

Beier, M. E., & Ackerman, P. L. (2005). Age, ability, and the role of prior knowledge on the acquisition of new domain knowledge: Promising results in a real-world learning environment. *Psychology and Aging*, *20*(2), 341-355.

Below, S. (2014, December 31). *New year, new workplace*! *SIOP announces top 10 workplace trends for 2015*. Retrieved from http://www.siop.org/article_view.aspx?article=1343

Berry, M. J. A., & Linoff, G. (1997). *Data mining techniques*. New York, NY: John Wiley.

Blok, A., & Pederson, M. A. (2014). Complimentary social science? Quali-quantitative experiments in a big data world. *Big Data & Society*, *1*, 1-6.

Botsford Morgan, W., Dunleavy, E., & DeVries, P. D. (2015). Using big data to create diversity and inclusion in organizations. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 310-335). New York, NY: Routledge.

Brieman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*, 199-231.

Burnham, K. P., & Anderson, D. R. (2004). *Model selection and multimodal inference: A practical information-theoretic approach*. New York, NY: Springer.

Campbell, J. J., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness*. New York, NY: McGraw-Hill.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., . . . Calatone, R. (2015). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*. Advance online publication. doi:10.1177/1094428115617004

Chung, C. K., & Pennebaker, J. W. (2012). *Counting little words in big data: The psychology of communities, cultures, and history. Social cognition and communication*. New York, NY: Psychology Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159. doi:10.1037/0033-2909.112.1.155

Collins, J. M., & Clark, M. R. (1993). An application of the theory of neural computation to the prediction of workplace behavior: An illustration and assessment of network analysis. *Personnel Psychology*, *46*, 503-524.

Colquitt, J. A., & Zapata-Phelan, C. P. (2007). Trends in theory building and theory testing: A five-decade study of the Academy of Management Journal. *Academy of Management Journal*, *50*, 1281-1303. doi:10.5465/AMJ.2007.28165855

Contractor, N. (2012, September). *Can big data motivate new theories and methods?* Opening keynote for the Digital Research 2012 Conference, Oxford, UK.

Derose, C. (2013, October 7). How Google uses data to build a better worker. *The Atlantic*. Retrieved from http://www.theatlantic.com/business/archive/2013/10/how-google-uses-data-to-build-a-better-worker/280347/

Dyche, J. (2012). Big data "Eurekas!" don't just happen. *Harvard Business Review*. Retrieved from https://hbr.org/2012/11/eureka-doesnt-just-happen/

Eagly, A. H., & Wood, W. (2013). The nature–nurture debates 25 years of challenges in understanding the psychology of gender. *Perspectives on Psychological Science*, *8*, 340-357.

Franks, B. (2012). *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics* (*Vol. 56*). New York, NY: John Wiley.

Freyer, D. (1930). The objective and subjective measurement of interests—An acceptance-rejection theory. *Journal of Applied Psychology*, *14*, 549-556. doi:10.1037/h0073774

Goldberg, A., Srivastava, S., Manian, V. G., & Potts, C. (2016, April). *Enculturation trajectories and individual attainment: An interactional language use model of cultural dynamics in organizations*. Paper presented at the Wharton People Analytics Conference, Philadelphia, PA.

Guastello, S. J., Koopmans, M., & Pincus, D. (Eds.). (2009). *Chaos and complexity in psychology: The theory of nonlinear dynamical systems*. New York, NY: Cambridge University Press.

Hambrick, D. C. (2007). The field of management's devotion to theory: Too much of a good thing? *Academy of Management Journal*, *50*, 1348-1352. doi:10.5465/AMJ.2007.28166119

Hernandez, I., Newman, D., & Jeon, G. (2016). Twitter analysis: Methods for data management and validation of a word count dictionary to measure city-level job satisfaction. In S. Tonidandel, E. King, & J. Cortina's (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 64-114). New York, NY: Routledge.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382-417.

Hsieh, N.-C., Hung, L.-P., & Ho, C.-H. (2009). A data driven ensemble classifier for credit scoring analysis. In T. Theeramunkong, B. Kijsirikul, N. Cercone, & T.-B. Ho (Eds.), *Advances in knowledge discovery and data mining* (pp.). Berlin, Germany: Springer-Verlag.

Ilies, R., Scott, B. A., & Judge, T. A. (2006). The interactive effects of personal traits and experienced states on intraindividual patterns of citizenship behavior. *Academy of Management Journal*, *49*, 561-575. doi:10.5465/AMJ.2006.21794672

Intrator, O., & Intrator, N. (2001). Interpreting neural-network results: A simulation study. *Computational Statistics & Data Analysis*, *37*, 373-393.

Jacobs, R., Murphy, K., & Silva, J. (2013). Unintended consequences of EEO enforcement policies: Being big is worse than being bad. *Journal of Business and Psychology*, *28*(4), 467-471.

Jarrett, R. P. (1918). A scale of intelligence of college students for the use of college appointment committees. *Journal of Applied Psychology*, *2*, 43-51. doi:10.1037/h0071002

Kaplan, D., & Chen, J. (2014). Bayesian model averaging for propensity score analysis. *Multivariate Behavioral Research*, *49*, 505-517.

Karanika-Murray, M., & Cox, T. (2010). The use of artificial neural networks and multiple linear regression in modelling work-health relationships: Translating theory into analytical practice. *European Journal of Work and Organizational Psychology*, *19*, 461-486.

Karasek, R. A., Jr. (1979). Job demands, job decision latitude, and mental strain: Implications for job redesign. *Administrative Science Quarterly*, *24*, 285-308.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196-217.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*, 1-12.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000081

Landis, R. S. (2014, May). Inductive reasoning: The promise of big data. In *Theme track: Deductive research meets inductive research*. Symposium conducted at the 29th annual meeting of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Landis, R. S., & Cortina, J. M. (2014). Is ours a hard science (and do we care)? In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 9-35). New York, NY: Routledge.

Laney, D. (2001). *3D data management: Controlling data volume, velocity and variety*. Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

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

Markon, K., & Chmielewski, M. (2013). The effect of response model misspecification and uncertainty on the psychometric properties of estimates. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (Vol. 66, pp. 85-114). Berlin, Germany: Springer.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356-376.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147-163. doi:10.1037/1082-989X.9.2.147

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537-563.

McKinley, J. C., & Hathaway, S. R. (1942). A Multiphasic Personality Schedule (Minnesota): IV. Psychasthenia. *Journal of Applied Psychology*, *26*, 614-624. doi:10.1037/h0063530

Muñoz, C., Smith, M., & Patil, D. J. (2016, May). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology*, *50*, 823-854.

Olguín, D. O., & Pentland, A. (2010, February). *Assessing group performance from collective behavior*. Paper presented at the 2010 Workshop on Collective Intelligence in Organizations: Toward a Research Agenda, Savannah, GA.

O'Neil, C. (2015, June 4). *Cathy O'Neil: Weapons of math destruction* (Video file). Retrieved from https://personaldemocracy.com/media/weapons-math-destruction

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169-198. doi:10.1613/jair.614.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., & Stillwell, D. J., . . . Seligman, M. E. P. (2014). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspp0000020

Pentland, S. (2012). The new science of building great teams. *Harvard Business Review*. Retrieved from https://hbr.org/2012/04/the-new-science-of-building-great-teams/ar/1

Pintner, R., & Toops, H. A. (1917). Mental tests of unemployed men. *Journal of Applied Psychology*, *1*, 325-341. doi:10.1037/h0074925

Plate, T. A., Bert, J., & Band, P. (2000). Visualizing the function computed by a feedforward neural network. *Neural Computation*, *12*, 1337-1353.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (in press). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*.

Putka, D. J., & Oswald, F. L. (2016). Implications of the big data movement for the advancement I-O science and practice. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp. 181-212). New York, NY: Routledge.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*, 1-39. doi:10.1007/s10462-009-9124-7

Savitz, E. (2013). Big data: Big hype? *Forbes*. Retrieved from http://www.forbes.com/sites/ciocentral/2013/02/04/big-data-big-hype/

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262.

Schmidt, H. O. (1945). Test profiles as a diagnostic aid: The Minnesota Multiphasic Inventory. *Journal of Applied Psychology*, *29*, 115-131. doi:10.1037/h0060192

Schmidt-Atzert, L., Krumm, S., & Lubbe, D. (2011). Toward stable predictions of apprentices' training success: Can artificial neural networks outperform linear predictions? *Journal of Personnel Psychology*, *10*, 34-42.

Schneider, B. (1987). The people make the place. *Personnel Psychology*, *40*, 437-453.

Seni, G., & Elder, J. (2010). *Ensemble methods in data mining: Improving accuracy through combining predictions*. San Rafael, CA: Morgan and Claypool.

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological Methods*, *18*, 572-582. doi:10.1037/a0034177

Somers, M. J. (1999). Application of two neural network paradigms to the study of voluntary employee turnover. *Journal of Applied Psychology*, *84*, 177-185.

Somers, M. J. (2001). Thinking differently: Assessing nonlinearities in the relationship between work attitudes and job performance using a Bayesian neural network. *Journal of Occupational and Organizational Psychology*, *74*, 47-61.

Somers, M. J., & Casal, J. C. (2009). Using artificial neural networks to model nonlinearity: The case of the job satisfaction-job performance relationship. *Organizational Research Methods*, *12*, 403-417.

Streitfeld, D. (2015, August 17). Data-crunching is coming to help your boss manage your time. *New York Times*.

Strong, E. J. (1951). Permanence of interest scores over 22 years. *Journal of Applied Psychology*, *35*, 89-91.

Strong, E. J. (1952). Nineteen-year followup of engineer interests. *Journal of Applied Psychology*, *36*, 65-74. doi:10.1037/h0056227

Sweeney, L. (2000). *Uniqueness of simple demographics in the U.S. population*. (Working Paper LIDAP-WP4). Pittsburgh, PA: Laboratory for International Data Privacy.

Tett, R. P., Walser, B., Brown, C., Simonet, D. V., & Tonidandel, S. (2013). The 2011 SIOP I-O Psychology Graduate Program Benchmarking Survey Part 3: Curriculum and competencies. *Industrial-Organizational Psychologist*, *50*(4), 69-90.

Tripathi, P., & Burleson, W. (2012, February). Predicting creativity in the wild: Experience sample and sociometric modeling of teams. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 1203-1212). New York, NY: ACM.

Ulrich, D. (2015, March 13). *Analyzing the analytics agenda by Dave Ulrich*. Retrieved from http://dialoguereview.com/analyzing-the-analytics-agenda-david-ulrich/

Waber, B. (2013). *People analytics: How social sensing technology will transform business and what it tells us about the future of work*. Upper Saddle River, NJ: Financial Times Press.

Waber, B. (2014, January 30). What data analytics says about gender inequality in the workplace. *Bloomberg BusinessWeek*. Retrieved from http://www.businessweek.com/articles/2014-01-30/gender-inequality-in-the-workplace-what-data-analytics-says

Welles, B. F. (2014). On minorities and outliers: The case for making big data small. *Big Data & Society*, *1*, 1-2.

Yan, J. L. S., McCracken, N., & Crowston, K. (2014a, March). *Semi-automatic content analysis of qualitative data*. Paper presented at iConference, Berlin, Germany.

Yan, J. L. S., McCracken, N., & Crowston, K. (2014b, June). *Design of an active learning system with human correction for content analysis*. Paper presented at the Workshop on Interactive Language Learning,

Visualization, and Interfaces, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD.

Yerkes, R. M. (1917). Psychology and national service. *Journal of Applied Psychology*, *1*, 301-304.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of the Sciences*, *112*, 1036-1040. doi:10.1073/pnas.1418680112

Author Biographies

Scott Tonidandel is the Wayne M. & Carolyn A. Watson Professor of Psychology at Davidson College and is a faculty affiliate of the Organizational Science PhD program at the University of North Carolina–Charlotte. He received his PhD in industrial-organizational psychology from Rice University. His recent work has focused on people analytics and the interface of big data and the organizational sciences. He coedited the SIOP Frontiers series volume titled *Big Data at Work: The Data Science Revolution and Organizational Psychology* and currently serves as associate editor for the *Journal of Business and Psychology* and *Organizational Research Methods*.

Eden B. King is an associate professor of psychology at George Mason University. She is pursuing a program of research that seeks to guide the equitable and effective management of diverse organizations. She is currently an associate editor for the *Journal of Management* and the *Journal of Business and Psychology* and is on the editorial board of the *Journal of Applied Psychology*.

Jose M. Cortina is professor in the I/O Psychology program at George Mason University. He received his PhD in 1994 from Michigan State University. His recent research has involved topics in meta-analysis, structural equation modeling, and philosophy of science. He has served previously as associate editor of the *Journal of Applied Psychology* and as editor of *Organizational Research Methods*.