ANNUAL REVIEWS

Annual Review of Organizational Psychology and Organizational Behavior

Big Data in Industrial-Organizational Psychology and Human Resource Management: Forward Progress for Organizational Research and Practice

Frederick L. Oswald,[1] Tara S. Behrend,[2] Dan J. Putka,[3] and Evan Sinar[4]

[1] Department of Psychological Sciences, Rice University, Houston, Texas 77005, USA; email: foswald@rice.edu

[2] Department of Organizational Sciences and Communication, George Washington University, Washington, DC 20052, USA

[3] Human Resources Research Organization, Alexandria, Virginia 22314, USA

[4] BetterUp, Pittsburgh, Pennsylvania 15243, USA

Annu. Rev. Organ. Psychol. Organ. Behav. 2020. 7:505–33

First published as a Review in Advance on November 5, 2019

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

https://doi.org/10.1146/annurev-orgpsych-032117-104553

Keywords

artificial intelligence, big data, personnel selection, talent management

Abstract

Big data and artificial intelligence (AI) have become quite compelling—and relevant, ideally—to organizations and the consulting services that help manage them. Researchers and practitioners in industrial-organizational psychology (IOP) and human resource management (HRM) can add significant value to big data and AI by offering their substantive expertise in how workforce-relevant data are measured and analyzed and how big data results are professionally, legally, and ethically interpreted and implemented by organizational decision makers, employees, policymakers, and other stakeholders in the employment arena. This article provides a perspective and framework for big data relevant to IOP and HRM that include both micro issues (e.g., linking data sources, decisions about which data to include, big data analytics) and macro issues (e.g., changing nature of big data, developing big data teams, educating professionals and graduate students, ethical and legal considerations). Ultimately, we strongly believe that IOP and HRM researchers and practitioners will become increasingly valuable for their contributions to the substance, technologies, algorithms, and communities that address big data, AI, and machine learning problems and applications in organizations relevant to their expertise.

INTRODUCTION

Whether you think about a virtual reality game that a job applicant might play, card readers that instantly record employee access to doors within an organization, or video monitoring of customer traffic in a store, big data pertaining to work and the workplace seem today to be as ubiquitous as the air we breathe. Sophisticated technologies for collecting and storing data allow for an exponential increase in organizational data that can be collected; and concomitant with (or perhaps because of) those technologies, sophisticated algorithms have emerged in parallel that can accommodate and analyze massive and messy data sets. With big data in hand, many complex research questions about employee and organizational phenomena can now begin to be addressed, if not definitively answered. Like you, we hold hope for the future.

Most articles addressing big data will begin with the question "What is big data?" In many research and practice arenas, the standard answer to this question usually contains several words starting with the letter V, including volume, variety, velocity, and veracity. In fact, we have located papers on big data with the number of Vs ranging from 3 to 42 (for the massive list of 42, see Shaffer 2017). Certainly, the Vs represent a useful mnemonic device for remembering the unique characteristics of large data sets that industrial-organizational psychology (IOP) and human resource management (HRM) researchers and practitioners must deal with as a set of unique opportunities and challenges. But however useful the Vs might be as words to describe big data, additional information is essential for serving organizations effectively.

In line with this latter point, our general approach to defining big data is to try to remain practical and problem-focused as a way to accumulate practical intelligence on big data questions from more of a bottom-up approach. Rather than thinking about the volume or velocity of organizational big data in the abstract, for example, an IOP or HRM researcher or practitioner might instead pose the more specific big data question, "How can I establish and test validation models that rely on big data from job applicant behaviors and job performance behaviors, both measured over time?" Or for someone dealing with a wide variety of big data, one might ask the specific question, "What is the most practical and useful way to merge and analyze intensive data (e.g., a daily mobile employee engagement survey) with extensive data (e.g., wide-ranging organizational data collected on job applicants and employees)?" In other words, it is through actual projects that deal with real-world challenges (e.g., data management, analytics, and ethics) that IOP and HRM researchers and practitioners can build up a broad and solid research and practice community that better understands the evolving meaning of big data and its analysis within the organizational and employment contexts.

With this philosophy in mind, the current review provides some important specifics regarding the landscape of big data in IOP and HRM, focusing on sources of data and the practical considerations for combining information across big data sources. This leads us to a discussion on the current and critical skills gap in data management and the analytical techniques most applicable for big data research, followed by proposed solutions. Next, we take a deeper look into

Table 1 Key factors for linking data sources

Factor	Description
Sources	Self-report surveys, mobile and biometric devices, and human resource information system records
Samples	Varying levels of analysis (individual, team, business unit, and organization)
Structure and consistency	Qualitative and quantitative data ranging from raw open-text data to well-audited financial data
Time intensity	Frequency of data collection, from multiple times per minute (customer interactions and purchases) to weekly or daily (team meetings) to once or twice a year (traditional performance reviews)
Country, language, and culture	Comparative equivalence of units of measurement, demographic categories, and item content
Timing and updating	Data across sources, gathered and updated asynchronously

the visualization, analytical, measurement, and technological aspects of the use of big data in IOP and HRM. Finally, after discussing what can be done technically with big data, we emphasize what should be done in terms of the ethical and legal considerations and guidelines that support the organizational research and practice community.

THE NATURE AND MANAGEMENT OF BIG DATA

Of all the adaptations associated with a big data approach to IOP and HRM, managing the data themselves is a major challenge—one that must be overcome before any analytics are even possible. In fact, "inaccurate, inconsistent, or hard-to-access data requiring too much manual manipulation" has often been cited as HR's biggest obstacle to make better use of data, metrics, and analytics (e.g., Harv. Bus. Rev. Anal. Serv. 2014). Data characteristics can introduce major, difficult, and occasionally unsurmountable challenges to the IOP or HRM research or practice professional engaging in big data analyses. It has been commonly reported that data cleaning—sometimes considered the "janitorial work" of data science (Lohr 2014)—is the most time-consuming phase of analytics work. A survey reported by Press (2016) cited cleaning and organizing data as comprising 60% of the work of data scientists, with an additional 19% spent collecting the data. Like the work of the janitor, data cleaning is a necessity of life that one should never take for granted. Ineffective data cleaning can lead to meaningless analyses and dangerous decision-making. IOP and HRM researchers should engage in effective data cleaning habits (and take pride in them).

Linking Data Sources

A critical issue underlying many big data challenges for IOP and HRM is the need to coordinate and connect data across widely varying sources (Ducey et al. 2015, King et al. 2016) (see **Table 1**). These sources include but are not limited to surveys, devices, and human resource information system (HRIS) records across many different levels of analysis, where levels reflect variations in (*a*) samples (e.g., individual, team, business unit, and organization-level data); (*b*) structure and consistency (e.g., ranging from raw open-text input to well-audited financial data); (*c*) time intensity (e.g., ranging from millions of customer transactions per day down to the traditional biannual performance review); (*d*) country, language, and cultural data that may require ensuring that the meaning of the data is equivalent after translation (e.g., item content, demographic categories, units of measurement and currency, and double-byte character systems); and (*e*) timing and updating, so there is knowledge of when each type of data happens or is predicted to be gathered and refreshed. Regarding this latter point, in contrast with a traditional research study where data collection happens once, it may be important for IOP and HRM researchers and practitioners to establish a continuous sustaining process to manage the variation in the arrival and updating of big data received over time.

As we know from traditional data analyses, unique identifiers must appear in every data set in order to be connected (as in a relational database) or merged together. For individuals, a permanent ID may be an employee ID number or corporate email address; however, as organizations increasingly plan their organizational analytics efforts in advance, unique ID assignments will be more common for each data collection effort. Unique IDs across data sets can then be crosswalked as the need arises; however, note that in this case, the file that crosswalks between these IDs needs to be highly secure, electronically encrypted, and protected (e.g., accessed by researchers on a limited project-driven basis). Note that in many settings, it can be impossible to connect IDs across data sets and data types. Or sometimes connections are possible but underappreciated (e.g., team-level data can be connected across sources or to individual-level data, so long as team membership is known). To the extent that missing and redundant IDs are found (a common encounter in our experience), and obviously, to the extent that flawed ID matches are made, this will limit the resulting accuracy, informativeness, and confidence in the use of data sets.

Which Variables Are Included in Big Data?

In traditional organizations and organizational research settings, data have tended to be collected deliberately, such as when job applicants are administered conscientiousness, job knowledge, and teamwork measures and their item-level scores are combined into their respective scale scores and used for personnel selection (or selection into future training once hired). These are relatively straightforward organizational situations and settings with which we are all familiar, where constructs and associated measures are prespecified. Even here, there are many challenges that professional standards of IOP and HRM research and practice have evolved to manage (e.g., SIOP 2018). Yet even with these challenges, this sets up an important contrast with situations in organizations where big data may be collected more incidentally and in real time, with perhaps less of a specific idea about how constructs of interest are being measured. Take a real-time electronic video recording of employees working in teams over the course of a year. What are all the constructs and job-related tasks that are being captured at the individual and team levels? They may not necessarily be obvious, nor may it be obvious how the measures of those constructs and tasks get defined, scored, and evaluated psychometrically, given the purposes of measurement (e.g., employee and team development, performance appraisal, or compensation).

Given unstructured big data like this (or other types of video, text, or audio), variables could still be identified top-down (e.g., trained subject matter experts could code video for team member behaviors that are preparatory, task based, interpersonal, adaptive, or regulatory in nature) (Rousseau et al. 2006). Variables can also be identified bottom-up from big data, such as through the algorithmic clustering of the data, and through interpreting patterns of relationships or networks between clusters. This bottom-up big data approach can be viewed as a modern version of the multitrait-multimethod matrix (Campbell & Fiske 1959) for assessing a nomological net of constructs (Cronbach & Meehl 1955).

Of course, as experienced IOP and HRM researchers, we are quite familiar with the research process, and yet key aspects of that process might often be neglected when working with big data. This is unfortunate, because whether one is collecting organizational big data through a top-down, bottom-up, or hybrid process with regard to the constructs or behaviors identified and measured, we can appreciate how accumulated knowledge from research and practice can systematically inform all aspects of this process. The types of measures and manipulations selected and implemented in a given organizational setting provide critical context around big data, in terms of understanding the data and the results that come from them. If theory is thought to play a role in these efforts, it is not as some abstract and monolithic idol to be worshipped, but rather as a set of systematic approaches informed by IOP and HRM to help address real-world organizational problems. In other words, theory is to be used as a tool and a means to an end; we need some theory as we work with big data, yet we should avoid theory for theory's sake (Campbell & Wilmot 2018).

The choice of whether and how to use data that are incidentally available in organizational databases but happen to be relevant to an organizational question or purpose at hand will most certainly involve (*a*) ethical and legal considerations involving data privacy concerns, (*b*) appropriate and ethical interpretation of results based on incidental data combined with black-box algorithms, and (*c*) the potential for revealing undesirable algorithmic regularities (e.g., so-called algorithmic biases) that are not tied to bona fide occupational qualifications of job applicants, business necessity, or other legitimate bases. Even in the era of big data and artificial intelligence (AI) (and perhaps especially so), IOP and HRM researchers and practitioners working in high-stakes contexts might lean toward identifying and incorporating measures that are construct relevant and associated with a dedicated purpose, because then their ethical, legal, and scientific justifications for the resulting data and analyses remain stronger.

To follow up on this sentiment, such an approach clearly does not prevent the spirit of big data and their analyses, where one incorporates measures that are more exploratory in nature (in other words, yes, we should continue to innovate with new measures in lower-stakes and research contexts), and it does not prevent big data algorithms being used to improve prediction (in other words, yes, we should still develop and implement AI and big data algorithms that are supported by robust out-of-sample predictions). But as soon as measures, samples, and settings change in a big data set, so might data-driven results—and the organizational decisions based on them (Whelan & DuVernet 2015). Therefore, in the era of big data, an organization should invoke the critical expertise of IOP and HRM experts—perhaps more so than ever before—to help ensure that big data, and the analyses and applications stemming from them, are scientifically, practically, ethically, and legally consistent and justifiable.

Assembling Big Data

As big data analytics are used extensively within an organization, HR, IOP, and HRM professionals often find themselves needing to bring separate organizational systems together, with each system using different conventions and data structures. Thus, the concept and challenge of integrating contained within siloed systems (Ryan & Herleman 2015) becomes increasingly relevant. In this data integration process, a series of critical steps is required (e.g., archive the original raw data; ensure the accuracy of ID matches and merges; make appropriate data-cleaning decisions; and during the process, record all steps that were taken for the sake of accuracy and future reference). Although detailed planning and partnerships assist in this process of ensuring the integrity of big data, opportunities for weak links and errors inevitably arise (Tonidandel et al. 2018). In some ways, this is no different than for traditional forms of data and analyses, but in other ways, new dimensions of consideration are involved.

When getting to the work of assembling HR big data sets, several related considerations are apparent, such as connecting files with appropriate identifiers, locating and reconciling redundant data between data sources, and ensuring that raw unstructured data are appropriately converted for analysis (e.g., text, audio, video). Because data are often gathered from different sources and for different purposes, it can be very helpful to explore and resolve their redundancies (across variables, but potentially across people or cases as well). Here, one could be addressing literal redundancies, such as having the same employee IDs and demographics recorded in several different data files (which can be challenging when the most complete data file is not necessarily the most accurate). There also may be close empirical redundancies, such as when two similar surveys of employee engagement are administered virtually simultaneously or when employee age and employee tenure are both in the data set but are highly correlated. Sometimes these redundancies mean deciding to omit or disregard variables or data sources if they appear to be redundant or poor proxies for other available data. Well-informed judgment is helpful before making such decisions, and IOP and HRM professionals are often essential to this end.

BIG DATA INFRASTRUCTURE

The increases in the Vs of big data in organizations (volume, variety, and otherwise) require that IOP and HRM researchers and practitioners, alongside their collaborators, gain a much deeper understanding of and capability for using the multiple systems that house and provision the big data necessary for analysis. These systems range from traditional HRIS to enterprise data warehouses, data lakes (i.e., raw data stored as any format or file type), and cloud-based platforms (Ryan & Herleman 2015), creating inevitable challenges when conducting organizational big data analyses, in terms of functionality and data formats (Angrave et al. 2016). Data integration may not be a simple matter, because certain systems and the data they house also have particular sensitivities and data-handling policies that require extreme caution for their use and storage, to avoid privacy, legal, and other ethical concerns (McLean et al. 2016).

Organizational big data analytics activities require IOP and HRM research teams to build and connect with knowledge of new content domains, such as programming, file structures, database terminology, application programming interfaces (APIs), and web scraping techniques (Braun et al. 2017). In addition, knowledge and application of computational modeling techniques allow HR analytics professionals to draw on and extend data gathered directly to simulate organizational phenomena (e.g., recruiting, selection, training, teamwork) and thereby meaningfully expand data-driven organizational decision-making capabilities (e.g., Kozlowski et al. 2016). Braun et al. (2017) extensively review many important concerns associated with big data analyses and also provide specific and useful guidance and recommendations. Particularly relevant to organizational big data are their discussions of identifying and obtaining data from database and web sources, computational requirements for big data processing, and special considerations involved in big data wrangling (e.g., managing big data files, ensuring tidy data, and dealing with missing and problematic data). Chen & Wojcik's (2016) guide to big data research in psychology is a similarly valuable resource for those engaged in organizational analytics, providing detailed discussions of data management planning, data acquisition, data processing, and data analytics in the context of psychological constructs.

BIG DATA SKILLS AND SKILLS GAPS

The Skills Gap

IOP and HRM professionals will often require additional training to obtain the necessary data gathering, data management, and data analytic skills that would leverage the potential of big data for improving their performance and effectiveness in an organization. For example, the "[l]ack of analytic acumen or skills among HR professionals" has been frequently cited as one of HR's biggest obstacles to the better use of business data, metrics, and analytics (Harv. Bus. Rev. Anal. Serv. 2014); similarly, a Society of Human Resource Management study (SHRM Found. 2016) reported that HR professionals were ill-equipped to use HR data for predicting workforce performance and improvement, conducting multiyear workforce planning, and correlating or otherwise modeling relationships between HR data and business performance (see also Angrave et al. 2016, Boudreau & Jesuthasan 2011).

Note that this latter problem of relating HR big data to organizational functions and outcomes is conceptually not much different than what prompted the historical development of utility analysis, which attempts to translate the validity of HR selection practices into metrics that managers understand, such as the dollar metric (e.g., Cascio et al. 2019). This big data skills gap is consequential for IOP and HRM researchers and practitioners gaining influence within an analytics-centric business environment and for avoiding the risks of being ignored or supplanted as an active participant in these discussions. Adhering solely to traditional statistical methods is often not an option in these new organizational contexts (Wax et al. 2015).

Addressing the Skills Gap

This big data skills gap has not gone unrecognized, with many organizations seeking to improve data management and analysis capabilities in these areas through a combination of recruiting and internal training (SHRM Found. 2016). The range of software used in the broader field of data science (which in turn is driving adoption of advanced analytics in organizations) includes those with which many IOP and HRM researchers are already becoming familiar, such as R and Python, as well as many other data management and analysis platforms, such as SQL, Hadoop, MapReduce, Unix Shell/AWK/Gawk, Apache Spark, and Java (see **http://www.datasciencecentral.com**, Ryan & Herleman 2015, Tonidandel et al. 2018). We realize that we may already be dating this article, as other new technologies take shape, but we want to risk being concrete about how many bases of specific knowledge and expertise are needed to coordinate big data–oriented organizational research.

And this relates to our next point: In developing big data analysis skills, IOP and HRM professionals alike must decide, given the nature of their relationships with organizations and value provided to them, how much to invest in specific technical knowledge and skills with particular programming languages and systems interfaces (which provide important specialized knowledge but have a slow learning curve), versus investing in broader general knowledge about big data analytics that inform intelligent decisions when collaborating with others who possess the specific technical knowledge and skills.

In many cases, the roles and projects for big data analytics in organizations may best be thought of in terms of a team or partnership model, in which no single individual possesses all the requisite skills to conduct an end-to-end project. Such teamwork addresses the reality of the sophistication of the work, the need for labor efficiency, and the need to keep institutional knowledge distributed (versus housed inside the brains of very few data gurus). Many organizations lack a dedicated big data analytics team (PWC 2015), or if they have one, the professions of their team members will span many functions, backgrounds, and specializations beyond IOP and HRM, such as mathematicians, labor economists, data scientists, market researchers, operations analysts, and visual designers (Bal 2016).

Within any data analytics team, we have experienced firsthand how professionals with advanced IOP and HRM degrees bring unique and valuable substantive knowledge about employment topics. Such professionals are encouraged to work, communicate, and listen vigorously, casting away any provincial mindset so they can work with and influence others effectively (Rotolo & Church 2015). Just as organizations benefit from data science, data scientists benefit from experts in organizational issues, such as those trained in IOP and HRM. Why? First, IOP and HRM experts are informed by evidence and experience from research and practice when they identify relevant ethical considerations. Their expertise, for example, can highlight procedural justice issues associated with communicating policies and practices around data collection, analysis, and sharing. Second, many IOP and HRM professionals are experts in ensuring that measures yield good data (not just big data) by psychometrically developing, validating, and otherwise vetting work-relevant predictors and criteria that actually measure the advertised constructs of interest. In support of this point, it can be useful to ask organizational leaders what conceptual and data-driven evidence supports a measure living up to its claims about measuring a construct, such as employee engagement or empathy, so that future discussion and investment in stronger measurement approaches might follow. Third, IOP and HRM professionals are experts in establishing research designs in the workplace that can credibly identify and isolate the effects of organizational practices and policies (Tonidandel et al. 2018), so then they can help ensure that big data analytics are not interpreted too aggressively or too narrowly for their practice and policy implications (Muñoz et al. 2016). Fourth, generalizing from the previous point, IOP and HRM professionals serve as a critical bridge between organizational decision makers and the organizational big data and machine learning tools that these decision makers earnestly seek out to address those problems. Conversations and collaborations between organizational roles and functions can thus be facilitated to understand how to develop an integrated organizational big data set, and then analyze and interpret it appropriately. As big data–related collaborations, training, and research efforts take hold organization-wide, the culture of the organization will likely change. In this sense, the big data movement within an organization alters the nature and scope of the big data that are actually collected.

We address the big data skills gap in IOP and HRM below by highlighting, in turn, the visualization, analytic, technological, and ethical dimensions of big data. We hope that readers will be informed by the article's general coverage, along with its emphasis on analytics, to then know where to go and learn more in this exciting arena, ultimately contributing their newfound knowledge and skills back to organizations and the IOP and HRM community.

BIG DATA VISUALIZATION

Although its roots lie in the oft-repeated (and oft-ignored) recommendation from introductory statistics and research methods courses to always plot and examine your data, big data pressures have pushed visualization techniques to new heights of relevance and utility. Early research (e.g., Al-Kassab et al. 2014) also supports visualization's incremental benefits for managerial decision-making beyond traditional forms of data presentation. Visualization is a way to explore and understand data and models interactively, dynamically, transparently, and intelligently, above and beyond what overall statistical summaries might provide. Specific examples of data visualization techniques are well matched to several of the aforementioned Vs of big data analytics: volume, velocity, variety, and veracity (Sinar 2015).

Regarding volume, visualization is an essential method for enabling IOP and HRM researchers, practitioners, decision makers, and other stakeholders to process and consolidate both the absolute and relative sizes and summaries of big data, based on the strategic placement of visual referents (e.g., comparing past and present characteristics of job applicants, job promotions, or regional sales). Regarding velocity, the speed at which HR big data arrives often precludes the use of manual processes to produce reviewable data. Visualizations can provide an audience with easily updated graphical displays and reporting to reduce information processing time and to speed the rate of decisions made based on HR data (Hans Rosling's Gapminder is one of many excellent examples; see **https://www.gapminder.org**). Regarding variety, the vast range of data pulled into HR analytics models can be aligned into common visual structures for joint display and interpretation (e.g., imagine 360-feedback information across hierarchies and networks of supervisors, teams, and employees). And finally, visualizations provide ways to directly examine and confirm data veracity by potentially allowing a researcher to identify out-of-range, interesting, or otherwise unusual data points.

Those using visualization should keep in mind two concerns that merit their own paragraph. The first concern is the well-known curse of dimensionality with big data (Domingos 2012). Data visualization might reveal interpretable patterns or outlying data points when certain variables or data summaries are involved, yet those visual discoveries might disappear under different yet equally reasonable combinations of variables or data summaries. The second concern is the important yet often neglected need to incorporate estimates of error into our visualization of big data and related findings. IOP and HRM researchers are accustomed to displaying error associated with bar charts and (occasionally) regression lines, but visualizing the errors associated with AI and machine learning algorithms associated with big data is a topic of current research (e.g., da Silva et al. 2017 have an interactive tool for visualizing the results from a random forests classifier).

Several contributors have bridged an extremely active and multidisciplinary literature on data visualization (almost wholly outside of IOP and HRM) with the application of these approaches in HR analytics. Sinar (2015) reviews the advantages of visualization as a communication tool for business audiences, given five common analytical topics: comparing categories, assessing hierarchies and part-to-whole relationships, showing changes over time, plotting connections and relationships, and mapping geospatial data. More recently, Tay and colleagues (2016) created a website (**http://www.graphicaldescriptives.org**), to enable social sciences researchers to input or load data and to readily create graphical plots associated with general descriptive statistics (mean, standard deviation, skew, and outliers), group mean differences, moderator analyses (e.g., bivariate scatter plots color coded by group membership), and the degree to which traditional statistical assumptions are met (e.g., linearity, normality, heteroscedasticity in multiple linear regression). Interactive visualization tools for big data and visualization tools for machine learning models applied to big data for clustering and prediction are also very important communication tools, and they continue to grow in their frequency of use (e.g., within R, see the use of plotly within ggplot2 at **https://plot.ly/ggplot2**).

BIG DATA ANALYSES AND ALGORITHMS

It is sometimes useful to remind ourselves of the big picture and overarching premises about conducting data analyses: Regardless of the amount of data that one has available, IOP and HRM researchers and practitioners conduct analyses in the attempt to satisfy many important organization- and employee-relevant goals. Analyses are performed to test hypotheses, evaluate theories, and ultimately gain insights into psychological phenomena in the workplace that are new, important, and generalizable. In organizational practice, analyses are performed to describe, classify, predict, and forecast organizational phenomena, often with the aim of making decisions about people, teams, and organizations that ultimately improve their functioning or well-being.

Organizations create the potential to reap substantial benefits through big data, even simply by tracking and describing basic HR characteristics of their respective workforces (e.g., percentage of successful hires, percentage of satisfied/engaged employees, and turnover rate), broken down by one or more elements of the organization's structure (e.g., division, business unit, region). Such descriptive statistics (and visualizations, as noted in the previous section) can be highly valuable and informative for directing attention to areas of the organization where intervention solutions may be needed (e.g., addressing recruiting, hiring, or performance management issues within units or regions). Yet obviously, simply describing aspects of an organization's workforce does little on its own to improve it. It is the quality and timeliness with which variables underlying those descriptions are measured, how IOP and HRM researchers and practitioners address these data, and how and when the organization acts on the analysis of that information that together determine their value.

Thus, many organizations will go beyond reporting simple descriptive statistics, conducting analyses aimed at classifying or clustering key organizational entities of interest, such as recruits, applicants, jobs, work units, or even survey and assessment items. For example, an organization may evaluate the roles or jobs in its workforce by clustering their similarity in terms of profiles of performance requirements or in terms of employee profiles of knowledge, skills, and abilities and other characteristics (KSAOs). The goal of such clustering might be to identify sets of related roles/jobs for purposes of streamlining and gaining efficiencies in recruiting, selection, workforce planning, and performance management efforts. This example involves clustering rows (e.g., people, jobs), but a more traditional connection with academic research comes in clustering columns (e.g., variables, items), such as when analyzing employee survey or assessment data using exploratory factor analysis. Such an analysis helps determine how survey items should be grouped together to form composites for reporting survey results and informing personnel decisions by summarizing the data in a clear, statistically principled, and effective manner.

The clustering analyses above lack a criterion to be predicted. By contrast, criterion-related validity analyses serve to explain or predict an organizational outcome (or outcomes) in the data set or extrapolate to forecast future events of a similar nature. For example, predicting turnover, counterproductive work behavior, or low employee engagement can be useful for targeting high-risk employees and delivering effective interventions to them. In addition, organizations may build models that predict business unit performance as some aggregated function of the performing individuals or teams that those units comprise, or even broader macro- and multilevel models that attempt to understand the impact that various HR and managerial interventions have on longer-term distal business outcomes (e.g., revenue by region, stock market performance).

EVOLUTION OF ANALYTIC METHODS, GRADUATE TRAINING, AND BIG DATA

Conducting organizational analyses requires specialized graduate training. To date, the training of graduate students in IOP and HRM has relied on a traditional set of statistical techniques to analyze organizational data. For example, graduate students in IOP are typically trained in the fundamentals of descriptive statistics, and they ideally would receive some exposure to psychometrics, factor analysis/clustering methods, experimental design, ANOVA, and simple variations on the generalized linear model (e.g., logistic regression, ordinary least squares regression) (Aiken et al. 2008, Tett et al. 2013), with an increased focus on structural equation modeling (SEM) and multilevel models (Aguinis et al. 2009, Austin et al. 2002). Graduate students in HRM might receive some of the aforementioned training, in addition to methodological training in other disciplines to which IOP is rarely exposed, such as econometrics, finance, and business strategy.

Thus, although graduate students in these fields appear to have a solid foundation for HR-related analyses in organizations, past academic researchers have also lamented that the time and training investment in graduate-level statistics and measurement seems to be waning in the field of psychology (Aiken et al. 2008, Merenda 2007).[1] Perhaps this decline in graduate training might reverse itself, given the broad multidisciplinary interest (and related grant funding) associated with addressing big data problems, technologies, and methods. That said, very few concrete applications of these methods to date appear in top-tier organizational research journals—but that appears to be changing. However, we have only to look to recent history to reflect on how Bayesian statistics

[1] Note that we know of no comparable examination of statistics and methodological education in management. What little evidence does exist suggests that IOP, HRM, and related graduate programs could all stand to offer additional statistics and methodological training (e.g., Tett et al. 2013).

has attracted multidisciplinary communities, yet there is relatively little training and participation from IOP and HRM (Zyphur et al. 2016). Perhaps a different story will emerge from graduate training in big data and machine learning, because organizations are themselves more involved in the analytic work. Overall, the benefit of participating in interdisciplinary methodological circles (big data, Bayesian, or otherwise), in addition to the obvious analytic benefits that might emerge, is to be able to create and sustain an expanded collaborative network of science and practice expertise directed at tackling large and challenging organizational problems.

Training IOP and HRM graduate students in machine learning has become an imperative, unless we seek to create another form of schism between organizational science and practice. Given over 20 years of machine learning developments that precede the awareness of our field, we find that the R programming language contains over 200 different machine learning methods available through R's caret package that perform classification/clustering, prediction, or both.[2] Furthermore, text data (e.g., open-ended interview data, job position descriptions) can be converted into meaningful quantitative information using the machine learning technique of natural language processing (NLP) (for a multitude of NLP methods, see **https://cran.r-project.org/web/views/NaturalLanguageProcessing.html**). Clearly, machine learning techniques applied to big data can complement and extend our very old friend, null hypothesis significance testing (NHST) based on the generalized linear model (ANOVA, linear regression).

Machine learning and AI techniques have made great strides in making predictions, as we can see across a range of games where humans now take second place: chess, Go, *Jeopardy!*, and poker. These techniques are now being developed beyond the domain of games and within the more complex and higher-stakes domains of medical diagnoses, self-driving cars, genetics, and geopolitics. Yet why have the staggering developments in big data been largely ignored in organizational research and practice? We posit that, in addition to the aforementioned needs for technical training, there is the fact that academic IOP and HRM training is situated more on the explainability side of an explainability-predictability continuum. In other words, machine learning and NLP methods arguably lend themselves better to inductive prediction models (without understanding which variables are necessarily responsible) rather than deductive process models of organizational phenomena (where some form of theory and constructs work together to influence the research sample, measures, and study design).[3] Theory-driven process models have been of primary interest in academic research; however, recent academic critiques have pointed out that an overemphasis on theory and theory development can constrain the value that can be obtained through inductive methods applied to the data gathering and data analysis process (e.g., Campbell & Wilmot 2018, Hambrick 2007, Locke 2007, Spector et al. 2014).

Notably, this relatively recent wave of encouragement to engage in induction largely ignores the machine learning clustering, text mining, and prediction approaches found in the big data domain. Yet the rich qualitative data used within inductive organizational research for decades bear important similarities to the rich data sources now gathered for bottom-up prediction using big data and machine learning techniques. It is as if the right brain is finally connecting with the left brain in the arena of organizational methodology and analysis.

Because the inductive-deductive cycle of research has gotten more diverse and complex, given so many forms of data and analysis, IOP and HRM researchers should take deliberate steps to plan and document their confirmatory and exploratory analyses to be conducted with organizational

[2] Caret is a machine learning package in R that offers a comprehensive framework for tuning, fitting, and cross-validating a wide variety of regression and classification models and algorithms that predict both continuous and categorical outcomes (**http://topepo.github.io/caret/index.html**).

[3] Putka et al. (2018), Sharpe (2013), and Tonidandel et al. (2018) offer additional perspectives on reasons for lack of dissemination of modern methods in organizational and psychological research.

data. Then, readers can be assured that they know how the research process was executed, can review the analysis code and reproduce reported results, know how the existing data (if shared) might be further explored, and know how research findings might be replicated and extended in future work. One might consider this transparency to be in the name of open science, but it is also a critical aspect of what we call open practice in dealing with all aspects of working with big data in organizations. Note that exploration and analysis of big data might often happen over time, in tandem with information systems that continue to provide those data. In these contexts, it might be especially important for IOP and HRM researchers and practitioners to follow the tenets of transparency, such as by documenting the confirmatory and exploratory aspects of one's research as it iterates, and by planning beforehand how and whether new incoming data will be combined with past data when generating new analyses.

WHAT IS REALLY NEW WITH BIG DATA ANALYSES?

The IOP and HRM disciplines largely apply traditional statistical methods (multiple linear regression, ANOVA, SEM, or mixed-effects models) that address a fairly circumscribed set of research designs. Published academic research has typically been plagued by small sample sizes and low statistical power, both being consistently lamented by the field for decades (e.g., Austin et al. 2002, Hunter & Schmidt 2015, Shen et al. 2011). Such research also has limited the number of variables involved in analyses, probably due to historical constraints on data collection (e.g., research conducted prior to the Internet and other technological innovations). Theories therefore have often remained relatively simple, given that a complex theory cannot be empirically supported based on research founded on small sample sizes and a limited number of variables.

In contrast, let us understand what is new in today's era of big data algorithms—both good and bad—by considering what happens when we free the constraints on traditional deductive theory and hypothesis testing found in published research. For instance:

- Instead of a small number of variables collected to test a theory of interest in a given study, what if we had access to far more variables for use in predictive and explanatory models that may be related to an organizational outcome of interest? Because such analyses are also based on very large samples, what if we discovered many empirically robust and predictive relationships as a result but they defied interpretation? How could we determine which of these predictive relationships might enhance prevailing theory or change management decisions?
- Instead of simply relying on traditional self-reporting (e.g., surveys, assessments) or other reporting measures (e.g., job performance ratings, interview ratings) that lend themselves to easy construction of quantitative scales that comprise the vast majority of our empirical research, what if we had access to substantial amounts of unstructured data—such as game data, text data from websites, and audio and video data from the workplace—for potential use in predictive and explanatory models relevant to organizations?
- Instead of using traditional statistical methods common to our field (e.g., moderated multiple regression, ANOVA), which cannot be used when the number of predictors exceeds the sample size, what big data methods can assist researchers in exploring and identifying complex nonlinearities and interactions when the number of predictors, in fact, might even exceed the sample size? Given massive amounts of data and associated statistical power, how might discoveries in this vein challenge traditional areas of HR research and practice?
- When the goal of a project or study is the prediction of an organizational outcome, what if we did not rely on a single model or a static sample? What if we instead generated multiple models that were fit to data obtained from bootstrapped samples (sampling randomly with

> replacement from the original sample)? Then we could average across the predictions made from multiple models for each case, in hopes of obtaining more robust and generalizable predictions. How would we deal with this approach to data analysis in the field of IOP and HRM, given our historical focus on testing relatively limited theory-driven models and interpreting their parameter estimates?

To be clear, there certainly remains a firm place for deductive research, with hypotheses to be developed and models to be tested a priori, based on sample sizes that provide adequate statistical power. But if IOP and HRM researchers and practitioners keep themselves situated within the research paradigm of NHST and confirmatory modeling, then we will rarely avail ourselves of the multidisciplinary engagement in big data opportunities and challenges, such as those raised above, that may also contribute to advancing both science and practice in organizations.

Interestingly, it is exactly in HR practice where one often finds a diverse array of big data, gathered for different purposes at different times, with unknown value for predicting and explaining outcomes of interest to our organizations. Academic researchers, as opposed to HR practitioners, are more likely to have much smaller sets of self-report measures administered for testing some prespecified theory or predicting some outcome. Thus, with big data, organizational practitioners should collaborate with academics from IOP and HRM to explore a large, potentially rich tapestry of structured and unstructured data to describe, classify, predict, and forecast various phenomena and outcomes of interest to organizations. One can dare to dream.

In the sections below, we offer several concrete big data examples from past work, and we discuss new analytic opportunities, often more inductive in nature, that arise in today's era of big data in organizations. In framing this discussion, we prioritize those areas that currently have the most potential for modern machine learning methods to advance science and practice beyond reliance on traditional statistical methods.

CAPITALIZING ON UNSTRUCTURED DATA

As one example, consider the processing and analysis of unstructured data such as text, graphics, or video. Such data may be extracted from social media or online sources such as Facebook, LinkedIn, or Twitter (e.g., Kosinski et al. 2016), or they may be scraped from various websites (e.g., Landers et al. 2016). Even traditional data sources can combine to form massive organizational data sets that link a range of unstructured data, such as open-ended responses as job applicants (e.g., employment application forms, employment interviews, assessment centers, resumes), with text-based workplace data (e.g., performance reviews, customer feedback surveys, employee and exit surveys), with employee descriptions of their changing work roles, and so on. Over the past few decades, developments in the areas of NLP algorithms and computing technology (software and hardware) have allowed one to take advantage of insights that come from efficiently converting unstructured linguistic data into structured data that are more amenable to analyses. This is similar to the tradition of recoding qualitative data into quantitative data, but on a vaster and more automated scale.

For example, Campion et al. (2016) used NLP algorithms, combined with simple linear regression models, to evaluate how well they could reproduce judges' evaluations of job applicants' written responses to accomplishment record questions that were focused on six competencies. They found their models could produce reasonable predictions, with cross-validated correlations ranging from $r = 0.59$ to 0.64, depending on the competency being assessed. These correlations were highly comparable to the estimates of reliability of human ratings based on a single judge obtained in their study.

Another specific example of relevant big data research lies in predicting the personality of users based on the text in their social media posts (e.g., Facebook, Twitter). For example, one study used NLP algorithms to process users' Facebook status messages and then used ridge regression (a linear regression method that accommodates big data) to predict these users' standing on a self-report measure of Big Five traits (Park et al. 2015). They found cross-validated correlations between the language-based personality trait estimates and self-reports of the Big Five to range from $r = 0.35$ (agreeableness and neuroticism) to 0.43 (openness to experience). Similarly, another study used Facebook "like" data in tandem with lasso regression models (another linear regression method; Tibshirani 1996) to predict self-reports of the Big Five (Youyou et al. 2015). Results indicated that for the average user (a Facebook user with 227 likes), Facebook "like" predictions had cross-validated correlations of $r = 0.56$ with user self-ratings across the Big Five traits. These estimates are in line with correlations found in the literature between ratings by self and spouse (0.58) and self and friend (0.45), and they were higher than correlations between self and work colleague (0.27), perhaps because of low acquaintance with some colleagues (see Connelly & Ones 2010). IBM's online Personality Insights application (**https://cloud.ibm.com/docs/services/personality-insights?topic=personality-insights-about**) and the University of Cambridge's Psychometrics Centre's online Apply Magic Sauce application (**http://applymagicsauce.com**) provide two examples of potential end products of this type of personality research that provide interesting illustrations—if not fair and legal implementation—of alternative organizational approaches to understanding personality and other stable individual differences characteristics of job applicants.

Moving to a video data example, researchers have offered a computational framework for automatically quantifying videotaped facial expressions, language, and speech patterns gathered in the course of employment interviews and using those data to predict interviewer ratings (Naim et al. 2016). This is one example of how phenomena and outcomes of interest to IOP and HRM researchers and practitioners might involve observations of human expressions, speech, and movements that extend well beyond traditional measurement and analysis methods (other examples follow). HR-oriented start-up companies are already selling products that combine video interviewing with prediction technology, under their general claim that better job applicants will be selected, and at a faster rate. More data and reliability and validity findings are needed to evaluate these claims, and questions certainly abound for the healthy skeptic (e.g., potentially confounding characteristics of applicants, such as age, race/ethnicity, and interview anxiety).

Where Can We Learn More?

Although they are not directly germane to the IOP and HRM fields, we suggest that readers review Kern et al. (2016), Kosinski et al. (2016), and Landers (2017) for very accessible introductions and tutorials on leveraging unstructured text data for use in modeling organizational phenomena of interest. Moreover, Landers et al. (2016) provide a clear tutorial geared toward psychologists on obtaining (scraping) data from websites. Although Kosinski et al. (2016) and Kern et al. (2016) focus on processing and analysis of social media text data, many of the methods they discuss have broad application to processing and analyzing other forms of unstructured text data. The Campion et al. (2016) and Park et al. (2015) studies in the previous section also provide specific, groundbreaking examples of publishing inductive NLP-based work in top-tier applied psychology journals.

HANDLING VERY HIGH *p*-TO-*N* RATIOS

Earlier, we mentioned that a typical big data situation is where the number of predictors may far exceed the sample size (i.e., $p > N$), and therefore traditional statistics cannot be applied to the data set. Park et al. (2015), in the work described above, had a model development sample of $N = 66{,}732$ and an initial set of predictors that numbered $p = 51{,}060$. Strictly speaking we do not have $p > N$ in this case; however, it is very close, and traditional linear regression analysis would be unable to analyze these data. In practical employment settings involving text data or in obtaining intensive repeated-measures data from wearable sensors (e.g., in studies of group/team communication or interaction patterns; Chaffin et al. 2017), p can far exceed N; this problem is virtually unavoidable in other research disciplines as well [e.g., neuroscience and functional magnetic resonance imaging (fMRI), or the analysis of microarray data in genetics].

The traditional statistical framework is model driven, meaning there is a strong need for p to be small and N large in the name of model parsimony, statistical power, and analytic tractability. Whenever p is large (e.g., hundreds of items across different measures), then in the traditional mode of analysis, we will usually employ data reduction methods, such as via simple unit-weighed sum scores, principal components analysis, or confirmatory factor analysis. Once the data are reduced to a manageable number of variables, then traditional analyses and theory testing take place.

As an alternative to the traditional methods above, one can instead address very high p-to-N ratios as part of the modeling process itself. For example, if one is dealing primarily with linear models, then one might adopt (*a*) lasso regression (Tibshirani 1996), which performs both predictor selection (i.e., many predictors in the big data set get zeroed out) and predictor weighting (i.e., the remaining predictors are assigned regression coefficients shrunk toward zero); (*b*) ridge regression, which retains all predictors and also shrinks their regression coefficients toward zero; or (*c*) elastic nets, which is a hybrid regression model that mathematically combines both lasso and ridge regression in a predictively optimal manner (Zou & Hastie 2005).

In the big data arena, one can also apply variations of the traditional data reduction techniques above that also incorporate the predicted outcome, when determining either the number of components to be extracted (e.g., principal components regression; Hastie et al. 2009) or the composition and number of those components (e.g., supervised principal and partial least squares regression; Bair et al. 2006, Lee et al. 2011). Large amounts of text data can involve data reduction techniques such as latent Dirichlet allocation (Blei et al. 2003) and randomized principal components analysis (Martinsson et al. 2011), which bear similarities to traditional cluster and factor analyses, respectively (Kosinski et al. 2016), and perhaps these text-reduction models can be extended to incorporate simultaneously the outcomes to be predicted.

In addition to handling the $p > N$ problem, it is safe to say that all predictive algorithms using big data attempt to find an optimal balance between the parsimony and complexity of the algorithms, given the central goal of maximizing cross-validated or out-of-sample model predictions. Striking this balance is critically important when p-to-N ratios are high, and doing so involves optimizing model tuning parameters (i.e., coefficients that help decide which model to choose out of a wide range tested) as well as model parameters themselves (i.e., coefficients for the model that was selected via the tuning parameters). Overall, this approach to predictive modeling is quite different from those for multiple linear regression, ANOVA, SEM, and other traditional statistical models.

Where Can We Learn More?

For more details on all the models mentioned above, we highly recommend books by Kuhn & Johnson (2013), which is more practitioner focused in providing R code examples throughout; James et al. (2013), which provides a useful blend of conceptual and theoretical description with practical implementation; and Hastie et al. (2009), which provides the most theoretical and mathematical treatment of these methods. These books remain highly used and useful, even with the passage of several years since their publication. Lastly, Putka et al. (2018) introduce lasso and elastic net regression models geared specifically toward organizational researchers, providing a concrete example of applying them to a leadership performance prediction problem.

IDENTIFYING AND ACCOUNTING FOR NONLINEARITY AND INTERACTIONS

In big data situations, the exact functional form of relationships between predictors and a criterion is often assumed to be unknown. Given very high *p*-to-*N* ratios and a lack of theory, detecting nonlinearities and interactions seems impossible, especially given the goal of finding a robust predictive model. The traditional tools of moderated multiple regression and polynomial regression will be impossible to implement in such situations. However, advances in predictive models over the past several decades provide alternatives for evaluating whether accounting for nonlinearities and interactions may offer any predictive or explanatory value above and beyond simple linear models. Although they are yet to be widely utilized in HR analytics, we briefly reference a sampling of these methods below.

With respect to examining nonlinearities, several methods represent simple variants on linear regression models. Examples include generalized additive models (Hastie & Tibshirani 1990), multivariate adaptive regression splines (MARS) (Friedman 1991), and hierarchical lasso (Bien et al. 2013). Although MARS and hierarchal lasso can help account for nonlinearities and interactions, their ability to explore higher-order interactions is limited. The next step up in complexity for exploring both nonlinearities and interactions involves methods that extend traditional classification and regression tree (CART) models (Breiman et al. 1984), such as random forests (Breiman 2001a) and stochastic gradient boosted trees (Friedman 2001, 2002; Putka et al. 2018, Sutton 2005). The final step up in complexity with respect to accounting for nonlinearities and interactions involves variations on support vector machines (Vapnik 1998) and modern variants on neural networks, namely deep learning (Goodfellow et al. 2016). Note that in random forests, support vector machines, and deep learning, stable complexities can be detected to benefit prediction without needing to model the underlying function that defines these complexities. This is exactly what distinguishes big data's algorithm-driven approach from the traditional model-driven approach (Breiman 2001b). There is a recent push toward explainable AI that attempts to make these black-box prediction methods more substantively interpretable (for a glimpse into these methods, see Ribiero et al. 2016).

Where Can We Learn More?

Beyond the general texts previously mentioned, which are very informative, Strobl et al. (2009) provide a tutorial on random forests geared toward psychologists, and Berk (2006) and Sutton (2005) provide very accessible introductions to random forests and gradient boosted trees. Elith et al. (2008) provide an exceptionally clear treatment of gradient boosted trees, although it is written from the perspective of the field of ecology. Miller et al. (2016) introduce a multivariate version of gradient boosted trees geared toward psychologists and apply the method to predict psychological well-being. Lastly, Oswald & Putka (2016, 2017) provide a nontechnical overview and concrete example of an application of random forests, stochastic gradient boosted trees, and support vector machines for predicting leader performance.

ACCOUNTING FOR AND BENEFITING FROM MODEL SELECTION UNCERTAINTY

All traditional inferential statistics involve estimating model parameters (e.g., means, correlations, regression coefficients, factor loadings) and their associated uncertainty due to sampling error variance. Sampling error variance represents the random idiosyncrasies of a sample, relative to the population of interest. Concerns over sampling error variance are highly apparent through extensive efforts on the part of IOP and HRM researchers to refine methods of meta-analysis and build up meta-analytic findings in various domains of study (Hunter & Schmidt 2015). Certainly, sampling error variance is an important source of uncertainty, but IOP and HRM researchers (like most social sciences researchers) have largely ignored uncertainty arising from the selection of the statistical model itself (Tonidandel et al. 2018). Models are abstractions, though, and therefore different statistical models can account for the same data, not just one. Chatfield succinctly summarized this point over 20 years ago:

> The estimation of model parameters traditionally assumes that the model has a prespecified known form and takes no account of possible uncertainty regarding model structure. In practice, model uncertainty is a fact of life and likely far more serious than other sources of uncertainty [e.g., sampling error variance] which have received far more attention from statisticians. This is true regardless of whether the model is specified on subject-matter [theoretical] grounds or, as is increasingly the case, when a model is formulated, fitted, and checked on the same data set in an iterative interactive way. (Chatfield 1995)

Model uncertainty is present regardless of whether one is dealing with simple regression models, multilevel models, SEM, or more sophisticated machine learning models based on big data. Model uncertainty and model equivalence have been important issues raised in the SEM literature (e.g., MacCallum et al. 1993, Preacher & Merkle 2012), where by focusing on just one model, IOP and HRM researchers may be ignoring inherent uncertainty in light of the theoretical and empirical reasonableness of alternative models. Additionally, organizational researchers might achieve new inductive insights by testing and exploring alternative models. Theoretically specifying a single model does not help one escape this aforementioned model uncertainty. In the framework of big data and complex algorithms, multiple models are entertained to balance model fit with model parsimony when clustering data or making predictions. Here, model uncertainty is inherent and made explicit.

Outside of IOP and HRM, the topic of model selection uncertainty in big data has been of growing interest, and it dovetails closely with the development of what are known as ensemble methods in machine learning (e.g., bagging, boosting, stacking, Bayesian model averaging) (Berk 2006, Hoeting et al. 1999). Essentially, ensembles are collections of predictions from different models or algorithms that are combined to make a new prediction; they provide a hedge against uncertainty associated with the perspective offered by any single model on its own. Model uncertainty becomes a strength, then, where not only can ensemble methods offer more robust predictions, but model uncertainty is also almost inherent to the methods for estimating associated prediction errors appropriately (e.g., Burnham & Anderson 2002). Although many ensemble methods are technical, two of the most publicly visible and straightforward applications of ensemble methods are not organizationally related but found in (*a*) hurricane forecasting, where popular weather outlets online will often average across hurricane paths different forecasting models with the aim of achieving more robust prediction; and (*b*) voting models, where ensembles that average across polls (along with their respective methodologies and biases) hope to yield more robust prediction, as has been found for predicting US presidential election outcomes (e.g., see Nate Silver's useful explanation at **https://goo.gl/avWa5N**).

In essence, one can draw a parallel between motivations behind model ensembles and meta-analyses. In meta-analyses, researchers' average effect sizes together (e.g., correlations) as a hopeful hedge against sampling error variance within and across different studies found in a research domain. Both meta-analysis and model ensembles are premised on the wisdom of the crowds notion that more generalizable estimates/predictions can be achieved by combining information from different sources of information (i.e., study estimates or modeling results).

Modern perspectives on model uncertainty suggest a new approach that extends the model-fitting approach we typically adopt in IOP and HRM research, one that involves leveraging information from multiple models to draw inferences regarding our parameters of interest and to make more robust predictions of outcomes of interest. Arguably, like the paths of hurricanes, human and organizational behaviors represent complex phenomena, and continuing to rely on single-model approaches and smaller data sets may be limiting (or misleading) our predictive and explanatory efforts in IOP and HRM fields in ways we may have yet to realize fully from big data and predictive modeling in the organizational setting.

Where Can We Learn More?

Work by Burnham & Anderson (2002) provides a very accessible introduction to the concept of model uncertainty in the context of linear regression models. They illustrate information theory–based methods for drawing inferences about regression parameters based not just on a single model but on multiple models that are weighted by the probability of being the closest-fitting model for the population of interest. Preacher & Merkle (2012) provide an overview of model uncertainty in SEM and methods for addressing it.

CONSIDERING THE PURPOSE FOR THE METHODS REVIEW

This section is an admittedly lengthy review of big data methods that pertain to IOP and HRM research and practice. We prioritize and provide it because most of the published works we found in this arena have been applied practice anecdotes that tend to be light on the specific examples and resources that we attempt to provide. We also tried to focus on those specific situations where modern methods can offer value beyond methods typically taught during graduate instruction in IOP and HRM. Finally, we made the earlier point that big data analyses are required when the number of variables or measures approaches the sample size—however, to be clear, big data analyses are also quite useful when the sample size remains much larger. The point of machine learning analyses is to mine data sets to detect stable-yet-complex relationships where they exist—as supported by the data under cross-validation. The following section describes a variety of measurement methods and technologies that yield the big data that afford this algorithmic possibility.

MEASUREMENT METHODS AND TECHNOLOGIES

Big data in the HR context can stem from virtually any source. Most of the examples below reflect innovative technological approaches to collecting big data, but as noted, even traditional HR data sources (e.g., job knowledge and personality measures, interview responses, and ratings data) may yield big data when collected over many people, settings, and time. It is an empirical question whether alternative or new sources of HR information will be superior to more traditional sources in terms of predicting organizational outcomes and informing HR decisions, and there is active debate about the legality of social media and other types of big data–related information being appropriated for personnel selection purposes (Chamorro-Premuzic et al. 2016). Keeping all this firmly in mind, several important sources of HR-relevant big data follow.

Serious Games and Gamification

Gamification refers to the incorporation of game-like elements in organizational contexts, and it is used frequently by organizations hoping to improve personnel selection; training; employee engagement; or some other job applicant, employee, or team experience. Game-like elements include the use of leaderboards to make relative standings salient, the awarding of badges or recognitions for achieving performance milestones, the establishment of goals or objectives, and the delivery of instant feedback and rewards based on goal progress. Landers (2014) notes that gamification tactics often result in rich data about user behavior but cautions that such tactics may not lead to intended outcomes (and of course, outcomes should be defined and not assumed).

For example, the game Airport Scanner, in which agents review various X-ray images and identify which ones are potentially hazardous, has important implications for the selection of security agents because it measures individual differences in visual search (Mitroff et al. 2015). Drawing here from a database that includes 3 billion trials from 12 million users, a wide range of highly powered statistical analyses might be impossible through other means. It is important to elaborate here on "garbage in, garbage out," however: A badly designed game will not be successful in attracting users, maintaining engagement, or, perhaps most importantly, providing construct-relevant or other job-relevant information. IOP and HRM researchers and practitioners who are looking to implement serious games within organizations are strongly encouraged to refer to professional test standards (SIOP 2018) and seek out the same critical information that they would for more traditional tests: for example, evidence that informs psychometric reliability, criterion-related validity, group mean differences (and adverse impact implications), and fairness. In the absence of such evidence, one is left with options such as trusting a test vendor's data-free claims or creating one's own organization-specific game without supportive evidence for its use. Both options are highly discouraged.

Internet of Things Device Data

Always-connected Internet devices, such as smart assistants, have emerged as a rich source of potential data. A variety of mobile and office technologies can recognize voice commands and execute various tasks for a user. That technology will "listen" continuously in case an employee issues a command, storing all voice data for later pattern analysis and generating a vast supply of organizational big data. Likewise, simple office devices, such as motion-detecting light switches and doors that detect employee access badges, can become a source of big data when activity is instantly recorded and stored via the cloud. Smart offices, equipped with these types of connected devices, could in theory provide multidimensional psychological data about employee behavior that inform performance management, training, and development, and even assist in the detection of illegal employee activity.

Cameras/Biometric Information

A variety of techniques have been developed for the analysis of video information, for use in applications such as automated on-site security and surveillance (Gandomi & Haider 2015), and for the generation of business intelligence, such as via targeted marketing. Robotic security guards are used in many offices to patrol using video surveillance, generating a detailed record of employees and other environmental features they detect on their routes, which can be analyzed for a variety of purposes. Cameras can also be attached to employees, such as in the case of police officers, delivery drivers, or warehouse workers who might wear body-mounted cameras, capturing real-time information about the world around them and about their own personal and interpersonal behaviors and reactions.

Social Media

The popular press has made the general public well aware that one can analyze big data from social media to inform decision-making—and sometimes unethically so. Examples of social media information relevant to IOP and HRM include the content of employee or organizational posts or blog content that could be used to generate insight about employee attitudes. Perhaps such insights could be shared privately and solely with the employee by the analysis tool itself, for developmental purposes. Career-related social media sites have proven to be a rich source of user data pertaining to the KSAOs that are required and desired for various jobs, where the analysis of these data could be viewed as next-generation job and occupational analysis. Many websites have researcher-friendly APIs that can be used to access data (see Chen & Wojcik 2016 for a detailed tutorial on accessing APIs from Reddit and Twitter). Social media sites such as Facebook and Twitter have proved to be useful tools for social scientists (e.g., Kosinski et al. 2015), but they clearly are not without serious considerations and controversy, inside and outside the personnel selection arena.

Text or Sentiment Analysis

Text data can be extracted from a wide range of HR sources, such as job applications, resume files, and interview recordings or transcripts. The analysis of text data can also take many approaches, including a user-defined dictionary approach, in which words that exist in a dictionary are analyzed for their frequency, categories, and co-occurrences. The dictionary approach is common across a broad range of psychological research fields, such as personality [e.g., the Linguistic Inquiry and Word Count (LIWC) dictionary; see Pennebaker et al. 2003]. Another approach is bottom-up in nature, where feature extraction and word co-occurrence analyses are extracted from patterns found in large bodies of text data, without a user-defined dictionary. A recent article in this journal provides a range of useful software options for text analysis (Short et al. 2018).

Mobile Sensors

Unobtrusive sensors can be embedded in ID tags, built into cell phone apps, or placed within other work-related devices, thus enabling the automatic collection of data about an employee's movements, physiology, proximity, interactions, and physiological states. Popular examples of this technology include Fitbit health sensors, radio-frequency identification–enabled ID tags, and sociometric badges. The last of these, sociometric badges (Olguín Olguín et al. 2009), are often worn like a pendant and are designed to capture a continuous and fine-grained stream of multimodal data pertaining to speech, movement, proximity, and group interaction patterns. Individual and team characteristics and dynamics can thus be captured unobtrusively and in real time along multiple psychological dimensions (e.g., performance, satisfaction, interpersonal communication, transfer of roles) (Kozlowski & Chao 2018).

Technological improvements of this nature, of course, do not guarantee better data: A series of well-designed studies conducted by Chaffin et al. (2017) illustrates both the "promise and perils" of wearable sensors as a data source for organizational research. Although these researchers found several instances of high data accuracy for wearable devices (most notably for gauging the colocation of multiple people), they also cautioned future researchers about the extensive planning, pretesting, and monitoring necessary to use these new data sources appropriately. On this latter point, sociometric badges and other high-intensity data-gathering devices can be affected by missing data due to signal loss (e.g., disconnection from a wireless Internet hub) and data entanglement (e.g., the same sensor picking up data from multiple individuals) (Braun et al. 2017). Employees might also retaliate or game the sensors (e.g., damage or deceive them) should monitoring be viewed as intrusive. Moral, ethical, and legal concerns should be top of mind for the IOP and HRM researcher and practitioner in these contexts, given that unobtrusive collection of high-intensity data not only can lead to employee anger and other forms of reactance and

Table 2 A sample of resources and additional information

Type of resource	Examples and references
Public data repositories	University of Chicago Workforce Data Initiative, **http://dsapp.uchicago.edu/resources/opensource**
	US government data, **http://data.gov**
	Open Science Framework, **http://osf.io**
	Open Psychological Datasets, **https://tinyurl.com/y2gyhns2**
	metaBUS, **http://metabus.org**
	Google Dataset Search, **https://toolbox.google.com/datasetsearch**
	Amazon Web Services Open Data, **https://registry.opendata.aws**
	Stanford Large Network Dataset Collection, **http://snap.stanford.edu/data**
Community forums	Data Science Central, **http://www.datasciencecentral.com**
	KDnuggets, **https://www.kdnuggets.com**
	Cross Validated, **https://stats.stackexchange.com**
	Stack Overflow, **https://stackoverflow.com/questions**
Visualization tools	dataviz.tools, **https://dataviz.tools**
	ggplot2, **https://r4ds.had.co.nz/data-visualisation.html**
	esquisse, **https://github.com/dreamRs/esquisse**
	Graphical Descriptives, **http://www.graphicaldescriptives.org**
Analysis and methodology tutorials	**Examples and references**
R programming	OpenIntro Statistics, **https://www.openintro.org/stat/labs.php?stat_lab_software=R**
	Data Science for Social Scientists, **http://datascience.tntlab.org/schedule-materials/**
	Applied Machine Learning Workshop at 2019 RStudio Conference, **https://github.com/topepo/rstudio-conf-2019/tree/master/Materials**
Unstructured text data	Landers 2017
	Kern et al. 2016
	Kosinski et al. 2016
	Short et al. 2018
Nonlinearity and interactions	Putka et al. 2018
	Strobl et al. 2009
	Berk 2006
	Sutton 2005
Model uncertainty	Preacher & Merkle 2012
	Burnham & Anderson 2002
Example publications	Campion et al. 2016
	Naim et al. 2016
	Park et al. 2015
	Youyou et al. 2015

negative performance consequences (Yost et al. 2018) but also raises legitimate concerns about privacy, informed consent, the need-to-know principle, and other contextual matters to which IO psychologists can speak, and big data and analytic methods themselves cannot.

Public Data Repositories

Many researchers and other institutions, many inspired by the open science movement, have made data sets freely available to the public for secondary analysis (see **Table 2**). Of special interest to us, the University of Chicago's Workforce Data Initiative is a source of data about skills, credentials, labor markets, and employment. Although primarily aimed at macrolevel researchers, the data sets are potentially valuable to IO psychologists and management scholars. Data.gov (**https://www.data.gov**), a project sponsored by the US government, hosts over 100 data sets that are relevant to IOP and HRM topics, such as the American Community Survey and the Job Openings and Turnover Survey. In terms of other useful public data repositories, the Open Science Framework (OSF; **http://osf.io**) is a fast-growing platform that hosts a wide array of research data sets and related study materials (e.g., experimental protocols, measures, programming code). The metaBUS platform (**http://metabus.org**) has coded and categorized over one million correlational effect sizes found in published organizational research papers spanning more than 10 years and 23 journals (Bosco et al. 2014). The platform provides an organizational research taxonomy to identify constructs, visualize the distribution of correlations between those constructs, and instantly meta-analyze those correlations. Finally, Google's relatively new database search engine (**https://toolbox.google.com/datasetsearch**) allows for readily locating publicly available databases, and one hopes additional data sets relevant to IOP and HRM researchers will accumulate.

Privately sponsored data sets are also available, such as those from Amazon Web Services' Public Data Sets repository (Chen & Wojcik 2016). The Stanford Large Network Dataset Collection database (**http://snap.stanford.edu/data**) can be used to generate insights about patterns of networking and collaboration. Note that in analyzing archival data sources, many complexities have long been identified, explored, and addressed (e.g., dealing with multilevel and missing data, complex sampling designs, and cross-cultural data) (see Trzesniewski et al. 2011). These lessons of experience, rather than being relearned, can be usefully applied to today's big data, even as big data analyses themselves continue to evolve.

Traditional Human Resource and Organizational Data

Finally, we note that personnel records about salary, performance, attendance, and tenure will remain valuable sources of information—perhaps even more so when they are combined with richer behavioral data. For example, turnover data are notoriously limited in their information value because specific information regarding why the employee left the organization is often not captured; usually only the binary outcome of leaving versus not leaving is available, perhaps along with a code for voluntary versus involuntary turnover. However, if organizations captured big data on employee interactions with teammates over time, or on employee satisfaction and engagement on a more frequent basis than the annual performance review, then our understanding and prediction of turnover might be enhanced above and beyond what is also usefully learned from individual differences and contextual correlates (e.g., Wanberg & Banas 2000).

The technological offerings of big data—those summarized above and others—should not be applied in the organizational setting without a measured and collaborative consideration of privacy, ethical, and legal concerns for all parties providing big data being harvested and analyzed (e.g., job applicants, employees, customers, clients, and organizations). Conversely, it might be an ethical oversight of one's professional responsibility not to do so. Some of these important concerns are reviewed in the next section.

PRIVACY, ETHICAL, AND LEGAL CONSIDERATIONS

Big data analyses in the HR and organizational settings introduce several unique ethical challenges relating to the protection of personal privacy and the provision of informed consent. Data privacy regulations can be enacted, such as the EU General Data Protection Regulation enacted in May 2018, but actual enforcement of such regulations in organizations requires resources for appropriate training and policies that will ultimately lead to demonstrating compliance (see

Tikkinen-Piri et al. 2018). Misuse of personal data by corporations and governmental entities has become an issue of great concern, and hundreds of news stories and books have detailed the many ways that big data can go wrong in HR and other organizational contexts.

For example, a heavily covered news story from 2016 involved the case of a health benefits management company using data based on employee habits and information obtained from fitness trackers to predict which groups of female employees from client companies might be more likely to become pregnant. Harrowingly, this information was then reported to the employing clients. In cases of this nature, two independent issues are of general concern. First, there is the risk of sensitive information being stolen by malicious actors, who could use it for blackmail, identity theft, or profit. Second, there is a nontrivial risk that employers, armed with this information, can use it to target certain employees unfairly—by firing those employees who are most likely to be more expensive (e.g., health care costs, leaves of absence). Although such firing would be illegal, it might also be difficult to prove.

As another example, advances in facial recognition research in the government and technology sectors have received considerable criticism regarding the ethical implications of collecting, storing, and sharing images of people without their consent. AI algorithms have recently become better at classifying human faces in nonideal realistic conditions such as poor lighting, occlusion by clothing, or partial images. Ethical responsibilities become blurred when one considers that many of the faces in the databases that trained the facial recognition algorithms were based on publicly available datasets. Employers and institutions can sell, and have sold, facial images in their databases without consent (Murgia 2019). The availability of such data, along with discussions about collecting such data in organizations, puts IOP and HRM researchers working in organizational contexts in a unique position to exercise their ethical professional responsibilities.

Ethical Codes and Guidelines

Psychologists in the IOP and HRM domains who are working in organizations are bound by two sets of ethical guidelines. First, they must adhere to the American Psychological Association's *Ethical Principles of Psychologists and Code of Conduct* (APA 2017), a professional code of ethics referred to below as the APA Code. Second, they are bound by a constellation of interrelated federal, state, and organizational legal requirements for the protection of employee data.

Regarding the latter, the US Department of Health and Human Services' Code of Federal Regulations Title 45 Part 46 (typically referred to as the Common Rule) dictates the protections that must be followed by federal agencies and academic institutions. But the foundation of the Common Rule was developed in a different era, and many psychologists therefore assert that it is not sufficient to provide guidance in today's era of big data and related analyses. For example, it may not be clear how far broad employee consent applies to archival organizational data sets that are used for IOP and HRM research purposes, let alone when those data sets are combined to form big data.

A similar situation holds for the original *Belmont Report* of 1978, which offers the central durable principles of respect for persons, beneficence, and justice; yet it is unclear whether modern institutional review board (IRB) regulations are sufficient to meet these objectives in practice. For example, a 2011 National Research Council panel recommended that studies be excluded from IRB review when they involve "benign interventions or interactions that are familiar in everyday life" (Fiske & Hauser 2014, p. 13675). However, neither IRBs nor organizations that might seek to adhere to similar core principles currently have much in the way of established, standardized, and effective methods or guidelines for assessing the potential for harm (or benignity) resulting from the collection, analysis, and interpretation of big data and the AI and machine learning tools that are applied. Employees may be anonymous within individual data sets but then become more identifiable to the extent that multiple data sets can be associated with one another and subjected to analyses that were not anticipated in each constituent data collection effort. Generally speaking, the IRB process and informed consent requirement of human subjects research does not yet appear to extend its reach very far into broader big-data-oriented research and application.

Legal Requirements

Legal regulations exist in many countries to protect personally identifiable information (PII). In the United States, the National Institute for Standards and Technology (NIST) offers guidelines for the protection of PII and lists examples such as name, Social Security number, facial image, and handwriting. Yet, as mentioned above, certain types of information may not appear to be sensitive on their own, yet they can still be used to identify individuals with surprising accuracy when combined with other data (Rocher et al. 2019); thus, a distinction can be made between identifying data and linkable data, where the latter refers to data that become identifying once related to or combined with other data. As a simple example, ZIP code and date of birth cannot identify a person's name and address on their own; however, identification may become possible for some people once the two pieces of data are combined.

The protection of PII becomes more complex in the context of combining data, and it cannot be accomplished by simply removing identifiers from the data. For instance, a person within a data set who is described by department, age, and gender might be sufficient to be identifiable or narrowed down to a small subset of people, depending on the size of the organization. Similarly, multiple disparate pieces of information gleaned from big data analyses can be used to triangulate on the identity of individuals (Berman 2013, Chen & Wojcik 2016). And text-based data can be analyzed for PII based on substance (e.g., education, job, and lifestyle) and content (e.g., complexity of vocabulary, word use, writing style).

Although it is often obvious to avoid intentional acts that will clearly result in harm from big data collection and analyses, unintended negative consequences might be difficult to anticipate. Among these is a notion similar to the Hawthorne studies of the 1920s, where employee video, audio, or other electronic surveillance in the service of big data might create social facilitation effects (Douthitt & Aiello 2001) that alter employee behavior for the better (e.g., by improving employee accountability) or for the worse (e.g., by eroding employee morale). As a second, more humorous example of unintended consequences, Kosinski et al. (2015) found that people who indicated they liked curly fries on their Facebook profile were more likely to score higher on a measure of intelligence. Once this finding was reported in the media, however, more Facebook users began to indicate that they too liked curly fries, thus undermining the initial big data research prediction (Chamorro-Premuzic et al. 2016). This illustrates an important principle: As soon as organizations and/or employees change their behavior in response to algorithmic findings, the algorithm itself may need to change to make effective future predictions.

If we assume that big data can be collected, analyzed, and protected under high-integrity research policies and codes of conduct, additional ethical questions also arise. Chief among these is the question of ownership: Who are the owners of big data, the collectors or the providers? Or should there be some mutual arrangement? Inside an organization, the ownership model may be murky as it is defined, enacted, and enforced, given that employment contracts may or may not specify terms that allow for data collection and surveillance activities.

Finally, it is important to consider explicitly the communication of data collection activities. Transparency about what is collected and for what purpose is essential to open science and open practice, to ensure that research is planful and conducted with integrity, and to increase the likelihood that employees will find the data collection to be understandable and acceptable. Perhaps we can imagine certain circumstances where employees of the future (not just employers) would even encourage such data collection, such as in cases where they clearly reaped the benefits of other prior big data–collection efforts in which they were involved.

CONCLUSION

This article has provided a state-of-the art overview of big data in HRM contexts from the multiple critical angles of data visualization, methodology, measurement and technologies, and ethics. Certainly, we hope the reader finds this article useful in whole or in part, and we wrote it with the hopes that any wisdom imparted will remain durable in light of a very rapidly evolving IOP and HRM big data community of science and practice, alongside the rapidly evolving technologies, data analytics, practices, and policies that will guide the future of big data as applied to organizations. Alongside this growth trajectory, we anticipate that big data and related analyses will also often become more integrated within any given organization, given that integrated enterprise systems are increasingly used to collect and coordinate data at the employee, team, and unit levels. All of this will be much more successful if organizations collaborate with talented IOP and HRM researchers who can effectively bridge across their disciplines and levels of analysis (Molloy et al. 2011) and thereby valuably inform all aspects of organizational big data efforts (e.g., data identification, collection, analysis, and interpretation). What technologies might yield HR big data and metrics, to be modeled for their real-time prediction of employee engagement, sales, customer satisfaction, product loyalty over time, and company reputation (SHRM Found. 2016)? How sensitive can we make such predictions, if we were to supplement big data analytics with more established theories and traditional theory-driven methods, along with data-informed computer simulations of the organizational system (e.g., changes in company financials, compensation, selection ratios, employee turnover, and teamwork; for the latter, see Grand et al. 2016)? Clearly, gathering big data in the attempt to capture the multidimensionality and dynamics of organizations, teams, and employees is both a blessing and in some ways a curse, but we firmly believe that the expertise of IOP and HRM scholars will greatly increase the likelihood of blessings. The future of big data in organizations likely holds many exciting possibilities, at least some of which no algorithms or humans will predict.

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

The authors wish to acknowledge Ahleah Miles for her diligent copyediting efforts, Marie-Thérèse Wright for her expert editorial support, and Herman Aguinis for his helpful editorial insights.

LITERATURE CITED

Aguinis H, Pierce CA, Bosco FA, Muslin IS. 2009. First decade of organizational research methods: trends in design, measurement, and data-analysis topics. *Organ. Res. Methods* 12:69–112

Aiken LS, West SG, Millsap RE. 2008. Graduate 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. *Am. Psychol.* 63:32–50

Al-Kassab J, Ouertani ZM, Schiuma G, Neely A. 2014. Information visualization to support management decisions. *Int. J. Inf. Technol. Decis. Mak.* 13:407–28

Angrave D, Charlwood A, Kirkpatrick I, Lawrence M, Stuart M. 2016. HR and analytics: why HR is set to fail the big data challenge. *Hum. Resour. Manag. J.* 26:1–11

APA (Am. Psychol. Assoc.). 2017. *Ethical Principles of Psychologists and Code of Conduct*. Washington, DC: Am. Psychol. Assoc.

Austin JT, Scherbaum CA, Mahlman RA. 2002. History of research methods in industrial and organizational psychology: measurement, design, analysis. In *Handbook of Research Methods in Industrial and Organizational Psychology*, ed. SG Rogelberg, pp. 1–33. Oxford, UK: Blackwell

Bair E, Hastie T, Debashis P, Tibshirani R. 2006. Prediction by supervised principal components. *J. Am. Stat. Assoc.* 101:119–37

Bal K. 2016. Building an HR analytics team. *Human Resource Executive*, March 16. **http://hrearchive.lrp.com/HRE/print.jhtml?id=534360037**

Berk RA. 2006. An introduction to ensemble methods for data analysis. *Sociol. Methods Res.* 34:263–95

Berman JJ. 2013. *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Waltham, MA: Morgan Kaufmann

Bien J, Taylor J, Tibshirani R. 2013. A lasso for hierarchal interactions. *Ann. Stat.* 41:1111–41

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022

Bosco FA, Uggerslev K, Steel P. 2014. Scientific findings as big data for research synthesis: the metaBUS Project. In *2014 IEEE International Conference on Big Data*, pp. 18–22. New York: IEEE

Boudreau JW, Jesuthasan R. 2011. *Transformative HR: How Great Companies Use Evidence-Based Change for Sustainable Advantage*. San Francisco, CA: Jossey-Bass

Braun MT, Kuljanin G, DeShon RP. 2017. Special considerations for the acquisition and wrangling of big data. *Organ. Res. Methods* 21:633–59

Breiman L. 2001a. Random forests. *Mach. Learn.* 45:5–32

Breiman L. 2001b. Statistical modeling: the two cultures. *Stat. Sci.* 16:199–215

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*. New York: Chapman & Hall

Burnham KP, Anderson DR. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer. 2nd ed.

Campbell DT, Fiske DW. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56(2):81–105

Campbell JP, Wilmot MP. 2018. The functioning of theory in IWOP. In *Handbook of Industrial, Work, and Organizational (IWOP) Psychology*, Vol. 1: *Personnel Psychology*, ed. N Anderson, DS Ones, HK Sinangil, C Viswesvaran, pp. 3–37. London: SAGE. 2nd ed.

Campion MC, Campion MA, Campion ED, Reider MH. 2016. Initial investigation into computer scoring of candidate essays for personnel selection. *J. Appl. Psychol.* 101:958–75

Cascio W, Boudreau J, Fink A. 2019. *Investing in People: Financial Impact of Human Resource Initiatives*. Alexandria, VA: Soc. Hum. Resour. Manag. 3rd ed.

Chaffin D, Heidl R, Hollenbeck JR, Howe M, Yu A, et al. 2017. The promise and perils of wearable sensors in organizational research. *Organ. Res. Methods* 20:3–31

Chamorro-Premuzic T, Winsborough D, Sherman RA, Hogan R. 2016. New talent signals: shiny new objects or a brave new world? *Ind. Organ. Psychol.* 9:621–40

Chatfield C. 1995. Model uncertainty, data mining, and statistical inference. *J. R. Stat. Soc. A* 158:419–66

Chen EE, Wojcik SP. 2016. A practical guide to big data research in psychology. *Psychol. Methods* 21:458–74

Connelly BS, Ones DS. 2010. An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. *Psychol. Bull.* 6:1092–122

Cronbach LJ, Meehl PE. 1955. Construct validity in psychological tests. *Psychol. Bull.* 52:281–302

da Silva N, Cook D, Lee E-K. 2017. Interactive graphics for visually diagnosing forest classifiers in R. arXiv:1704.02502 [stat.ML]

Domingos P. 2012. A few useful things to know about machine learning. *Commun. ACM* 55:78–87

Douthitt EA, Aiello JR. 2001. The role of participation and control in the effects of computer monitoring on fairness perceptions, task satisfaction, and performance. *J. Appl. Psychol.* 86:867–74

Ducey AJ, Guenole N, Weiner SP, Herleman HA, Gibby RE, Delany T. 2015. I-Os in the vanguard of big data analytics and privacy. *Ind. Organ. Psychol.* 8:555–63

Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77:802–13

Fiske ST, Hauser RM. 2014. Protecting human research participants in the age of big data. *PNAS* 111(38):13675–76

Friedman J. 1991. Multivariate adaptive regression splines. *Ann. Stat.* 19:1–67

Friedman J. 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29:1189–232

Friedman J. 2002. Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38:367–78

Gandomi A, Haider M. 2015. Beyond the hype: big data concepts, methods, and analytics. *Int. J. Inf. Manag.* 35:137–44

Goodfellow I, Bengio Y, Courville A. 2016. *Deep Learning*. Cambridge, MA: MIT Press

Grand JA, Braun MT, Kuljanin G, Kozlowski SWJ, Chao GT. 2016. The dynamics of team cognition: a process-oriented theory of knowledge emergence in teams. *J. Appl. Psychol.* 101:1353–85

Hambrick DC. 2007. The field of management's devotion to theory: too much of a good thing? *Acad. Manag. J.* 50:1346–52

Harv. Bus. Rev. Anal. Serv. 2014. *HR joins the analytics revolution*. Rep., Harvard Bus. Sch. Publ., Boston

Hastie T, Tibshirani R. 1990. *Generalized Additive Models*. London: Chapman and Hall

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. 2nd ed.

Hoeting JA, Madigan D, Raferty AE, Volinksy CT. 1999. Bayesian model averaging: a tutorial. *Stat. Sci.* 14:382–417

Hunter JE, Schmidt FL. 2015. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: SAGE. 3rd ed.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer

Kern ML, Park G, Eischstaedt JC, Schwartz HA, Sap M, et al. 2016. Gaining insights from social media language: methodologies and challenges. *Psychol. Methods* 21:507–25

King EB, Tonidandel S, Cortina JM, Fink AA. 2016. Building understanding of the data science revolution and IO psychology. In *Big Data at Work: The Data Science Revolution and Organizational Psychology*, ed. S Tonidandel, EB King, JM Cortina, pp. 1–15. New York: Routledge

Kosinski M, Matz SC, Gosling SD, Popov V, Stillwell D. 2015. Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. *Am. Psychol.* 70:543–56

Kosinski M, Wang Y, Lakkaraju H, Leskovec J. 2016. Mining big data to extract patterns and predict real-life outcomes. *Psychol. Methods* 21:493–506

Kozlowski SWJ, Chao GT. 2018. Unpacking team process dynamics and emergent phenomena: challenges, conceptual advances, and innovative methods. *Am. Psychol.* 73:576–92

Kozlowski SWJ, Chao GT, Grand JA, Braun MT, Kuljanin G. 2016. Capturing the multilevel dynamics of emergence: computational modeling, simulation, and virtual experimentation. *Organ. Psychol. Rev.* 6:3–33

Kuhn M, Johnson K. 2013. *Applied Predictive Modeling*. New York: Springer

Landers RN. 2014. Developing a theory of gamified learning: linking serious games and gamification of learning. *Simul. Gaming* 45:752–68

Landers RN. 2017. A crash course in natural language processing. *Ind. Psychol.* 54:5–16

Landers RN, Brusso RC, Cavanaugh KJ, Collmus AB. 2016. A primer on theory-driven web-scraping: automatic extraction of big data from the Internet for use in psychological research. *Psychol. Methods* 21:475–92

Lee D, Lee W, Lee Y, Pawitan Y. 2011. Sparse partial least-squares regression and its applications to high-throughput data analysis. *Chemom. Intell. Lab. Syst.* 109:1–8

Locke EA. 2007. The case for inductive theory building. *J. Manag.* 33:867–90

Lohr S. 2014. For big-data scientists, 'janitor work' is key hurdle to insights. *New York Times*, Aug. 17. **https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html**

MacCallum RC, Wegener DT, Uchino BN, Fabrigar LR. 1993. The problem of equivalent models in applications of covariance structure analysis. *Psychol. Bull.* 114:185–99

Martinsson PG, Rokhlin V, Tygert M. 2011. A randomized algorithm for the decomposition of matrices. *Appl. Comput. Harmonic Anal.* 30:47–68

McLean S, Stakim C, Timner H, Lyon C. 2016. Big data and human resources—letting the computer decide? *Scitech Lawyer* 12:20–23

Merenda PF. 2007. Psychometrics and psychometricians in the 20th and 21st centuries: how it was in the 20th century and how it is now. *Percept. Motor Skills* 104:3–20

Miller PJ, Lubke GH, McArtor DB, Bergeman CS. 2016. Finding structure in data using multivariate tree boosting. *Psychol. Methods* 21:583–602

Mitroff SR, Biggs AT, Adamo SH, Dowd EW, Winkle J, Clark K. 2015. What can 1 billion trials tell us about visual search? *J. Exp. Psychol. Hum. Percept. Perform.* 41:1–5

Molloy JC, Ployhart RE, Wright PM. 2011. The myth of "the" micro-macro divide: bridging system-level and disciplinary divides. *J. Manag.* 37:581–609

Muñoz C, Smith M, Patil DJ. 2016. *Big data: a report on algorithmic systems, opportunity, and civil rights*. Rep., Exec. Office Pres., Washington, DC. **https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf**

Murgia M. 2019. Who's using your face? The ugly truth about facial recognition. *Financial Times*. April 19. **https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e**

Naim I, Tanveer MI, Gildea D, Hoque ME. 2016. Automated analysis and prediction of job interview performance. *IEEE Trans. Affect. Comput.* 99:191–204

Olguín Olguín D, Waber BN, Kim T, Mohan A, Ara K, Pentland A. 2009. Sensible organizations: technology and methodology for automatically measuring organizational behavior. *IEEE Trans. Syst. Man Cybern. B Cybern.* 39:43–55

Oswald FL, Putka DJ. 2016. Statistical methods for big data: a scenic tour. In *Big Data at Work: The Data Science Revolution and Organizational Psychology*, ed. S Tonidandel, EB King, JM Cortina, pp. 43–63. New York: Routledge

Oswald FL, Putka DJ. 2017. Big data methods in the social sciences. *Curr. Opin. Behav. Sci.* 18:103–6

Park G, Schwartz HA, Eichstaedt JC, Kern ML, Kosinski M, et al. 2015. Automatic personality assessment through social media language. *J. Personal. Soc. Psychol.* 108:934–52

Pennebaker JW, Mehl MR, Niederhoffer KG. 2003. Psychological aspects of natural language use: our words, our selves. *Annu. Rev. Psychol.* 54:547–77

Preacher KJ, Merkle EC. 2012. The problem of model selection uncertainty in structural equation modeling. *Psychol. Methods* 17:1–14

Press G. 2016. Cleaning big data: most time-consuming, least enjoyable data science task, survey says. *Fortune*. March 23. **https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#7b9080bc6f63**

PriceWaterhouseCoopers. 2015. *Innovating for tomorrow's workforce: transformation enabled by HR in the Cloud*. Survey, PriceWaterhouseCoopers, London. **https://roubler.com/au/wp-content/uploads/sites/9/2017/02/hr-tech-survey.pdf**

Putka DJ, Beatty AS, Reeder MC. 2018. Modern prediction methods: new perspectives on a common problem. *Organ. Res. Methods* 21:689–732

Ribiero MT, Singh S, Guestrin C. 2016. Local interpretable model-agnostic explanations (LIME): an introduction. *O'Reilly*. Aug. 12. **https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime**

Rocher L, Hendrickx JM, de Montjoye Y-A. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* 10:3069

Rotolo CT, Church AH. 2015. Big data recommendations for industrial-organizational psychology: Are we in Whoville? *Ind. Organ. Psychol.* 8:515–20

Rousseau V, Aubé C, Savoie A. 2006. Teamwork behaviors: a review and integration of frameworks. *Small Group Res.* 37:540–70

Ryan J, Herleman H. 2015. A big data platform for workforce analytics. In *Big Data at Work: The Data Science Revolution and Organizational Psychology*, ed. S Tonidandel, EB King, JM Cortina, pp. 19–42. New York: Routledge

Shaffer T. 2017. The 42 V's of big data and data science. *KDnuggets*, April. **https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html**

Sharpe D. 2013. Why the resistance to statistical innovations? Bridging the communication gap. *Psychol. Methods* 18:572–82

Shen W, Kiger TB, Davies SE, Rasch RL, Simon KM, Ones DS. 2011. Samples in applied psychology: over a decade of research in review. *J. Appl. Psychol.* 96:1055–64

Short JC, McKenny AF, Reid SW. 2018. More than words? Computer-aided text analysis in organizational behavior and psychology research. *Annu. Rev. Organ. Psychol. Organ. Behav.* 5:415–35

SHRM (Soc. Hum. Resour. Manag.) Found. 2016. *Use of Workforce Analytics for Competitive Advantage*. New York: The Econ. Intell. Unit

Sinar EF. 2015. Data visualization. In *Big Data at Work: The Data Science Revolution and Organizational Psychology*, ed. S Tonidandel, EB King, JM Cortina, pp. 115–57. New York: Routledge

SIOP (Soc. Ind. Organ. Psychol.). 2018. Principles for the validation and use of personnel selection procedures. *Ind. Organ. Psychol.* 11(Suppl. S1):1–97

Spector PE, Rogelberg SG, Ryan AM, Schmitt N, Zedeck S. 2014. Moving the pendulum back to the middle: reflections on and introduction to the inductive research special issue of *Journal of Business and Psychology*. *J. Bus. Psychol.* 29:499–502

Strobl C, Malley J, Tutz G. 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 14:323–48

Sutton CD. 2005. Classification and regression trees, bagging, and boosting. *Handb. Stat.* 24:303–29

Tay L, Parrigon S, Huang Q, LeBreton JM. 2016. Graphical descriptives: a way to improve data transparency and methodological rigor in psychology. *Perspect. Psychol. Sci.* 11(5):692–701

Tett RP, Walser B, Brown C, Simonet DV, Tonidandel S. 2013. The 2011 SIOP I-O Psychology Graduate Program Benchmarking Survey Part 3: curriculum and competencies. *Ind.-Organ. Psychol.* 50:69–89

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58:267–88

Tikkinen-Piri C, Rohunen A, Markkula J. 2018. EU General Data Protection Regulation: changes and implications for personal data collecting companies. *Comput. Law Secur. Rev.* 34:134–53

Tonidandel S, King EB, Cortina JM. 2018. Big data methods: leveraging modern data analytic techniques to build organizational science. *Organ. Res. Methods* 21:525–47

Trzesniewski KH, Donnellan MB, Lucas RE. 2011. *Secondary Data Analysis: An Introduction for Psychologists*. Washington, DC: Am. Psychol. Assoc.

Vapnik V. 1998. *Statistical Learning Theory*. New York: Wiley & Sons

Wanberg CR, Banas JT. 2000. Predictors and outcomes of openness to changes in a reorganizing workplace. *J. Appl. Psychol.* 85:132–42

Wax A, Asencio R, Carter DR. 2015. Thinking big about big data. *Ind. Organ. Psychol.* 8:545–50

Whelan TJ, DuVernet AM. 2015. The big duplicity of big data. *Ind. Organ. Psychol.* 8:509–15

Yost A, Behrend TS, Howardson GN, Darrow JB, Jensen J. 2018. Reactance to electronic surveillance: a test of antecedents and outcomes. *J. Bus. Psychol.* 34:71–86

Youyou W, Kosinski M, Stillwell D. 2015. Computer-based personality judgments are more accurate than those made by humans. *PNAS* 112:1036–40

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67:301–20

Zyphur MJ, Oswald FL, Rupp DE. 2016. Rendezvous overdue: Bayes analysis meets organizational research. *J. Manag.* 41:387–89