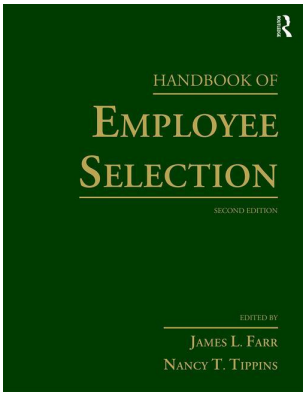

Handbook of Employee Selection

James L. Farr, Nancy T. Tippins, Walter C. Borman, David Chan, Michael D. Coovert, Rick Jacobs, P. Richard Jeanneret, Jerard F. Kehoe, Filip Lievens, S. Morton McPhail, Kevin R. Murphy, Robert E. Ployhart, Elaine D. Pulakos, Douglas H. Reynolds, Ann Marie Ryan, Neal Schmitt, Benjamin Schneider

Using Big Data to Enhance Staffing

Publication details
https://test.routledgehandbooks.com/doi/10.4324/9781315690193-43
Richard N. Landers, Alexis A. Fink, Andrew B. Collmus
Published online on: 22 Mar 2017

43

Using Big Data to Enhance Staffing

Vast Untapped Resources or Tempting Honeypot?[1]

RICHARD N. LANDERS, ALEXIS A. FINK, AND ANDREW B. COLLMUS

The overall purpose of organizational staffing is to deliver fresh hires into organizations. Efforts to improve staffing have historically involved pursuing two primary goals: improving job applicant quality and improving the process used to quantify and make decisions about those applicants. Industrial/-organizational (I-O) psychologists, based upon decades of research, have many specific processes they commonly employ to meet these goals. Despite this, a family of technologies commonly referred to as big data has begun to appear in staffing processes without much, if any, validation from I-O psychologists. Data scientists have claimed that such technologies have the potential to "disrupt" the bedrock staffing procedures on which much of modern I-O psychology has been built. The truth of this claim is difficult to determine for many reasons, but most glaringly because data scientists and I-O psychologists come from such different theoretical perspectives that it is often difficult to find common ground even in casual conversation.

As noted above, I-O psychologists rely upon a great deal of existing research to support the consideration of a wide range of individual differences as predictors in selection systems (i.e., KSAOs: knowledge, skills, abilities, and other characteristics) alongside methods to measure them (e.g., surveys and questionnaires, assessment centers, work sample tests). Using this toolkit, I-O psychologists can consistently improve hiring outcomes in terms of applicant reactions, task performance, and/or contextual performance for just about any organization. Importantly, the development of this approach was based upon a number of assumptions and theoretical perspectives that are not shared by everyone attempting to improve staffing. Specifically, I-O psychologists primarily practice from behind the broader assumptions of psychological science and the measurement guidelines commonly associated with it. Our science is one of theory and reflective constructs; that is, we assume certain persistent underlying human characteristics exist regardless of our measurement of them and that the data we solicit from job applicants are reflections of those characteristics. These assumptions are driven by psychological theory that was created, developed, and refined by psychological researchers based upon the scientific method over many decades, if not longer. For example, we have theory to suggest that there are persistent non-cognitive differences between people, which we call personality, and that these differences are associated with work-related outcomes, including job performance (Barrick & Mount, 1991). Thus we might administer a personality measure to job applicants as part of a selection system in order to predict their future job performance.

Scientists in other fields of inquiry that are highly relevant to staffing do not necessarily share these same values. The empirical branch of computer science, for example, is primarily

concerned with the development and testing of computers and related technologies, including algorithms (Newell & Simon, 1976). From this perspective, the world of computer languages is much more "real" than that of psychological constructs; there is no unmeasurable, unknowable characteristic of a computer that must be assumed to exist, tested only by proxy and by inference. To many computer scientists, a psychological construct is itself inherently unknowable, and, taken to its logical conclusion, studying the unknowable is a waste of researcher time and effort. In contrast, the patterns within data potentially caused by such constructs are a well-defined problem. They are data, and patterns with data can be modelled. With sufficiently high-quality data, such models could be used to predict other data that do not yet exist. One never needs to worry about constructs; the patterns tell the story. Thus, the major objectives for computer scientists in this domain are to increase the quantity of data from which to create models and to improve the predictive value of modelling. This sort of thinking lies at the foundation of big data and, to a degree, at the foundation of the older and broader field of business intelligence/business analytics (Chen, Chiang, & Storey, 2012). From this perspective, data are not necessarily reflective of a larger problem to be solved; they *are* the problem to be solved.

Until recently, practitioners of business analytics and big data analytics have applied this perspective primarily as a means to increase the effectiveness of marketing. For example, to maximize conversion from website visitor and online advertisement-viewer to purchaser of products and services, marketers collect and interpret incredibly vast sources of behavioral data as they occur. Computer clicks and taps, the keywords and phrases uses in search engines, specific webpages visited and the amount of time spent on them, the roads a smartphone has travelled down during a person's commute, the advertisements and conversations that a smartphone has overheard throughout the day, and many other such sources of information may all be collected and tied to a particular digital identity. An algorithm, refined automatically from these vast data sets that are continually updated to maximize prediction, is used to automatically identify advertisement content that maximizes click-through rates and displays advertising content when online shopping, all in a fraction of a second. This algorithmic approach has become so effective that many consumers view the accuracy of such systems as emotionally disturbing (Ur, Leon, Cranor, Shay, & Wang, 2012), due in part to the significant number of perceived privacy violations (Cumbley & Church, 2013). Regardless of the ethics of these practices, the sheer amount of detailed information available about almost everyone with access to the Internet has grown exponentially. As these data sources have grown in size and complexity, researchers have continued to expand the analytic toolkits used to make sense of and draw conclusions from them.

Because both these data sources and analytic toolkits offer a great deal of potential for staffing, the purpose of this chapter is to explore how this potential might be realized and how researchers and practitioners are already realizing it. Perhaps more importantly, we also explore *if* it should be realized. Big data are not necessarily high-quality data, and I-O psychology already has many techniques to obtain, analyze, and apply high-quality small data. Research is not yet available demonstrating specific validity or utility advantages to big data staffing approaches above and beyond more well-established small data techniques, and ultimately, big data may be little more than a fad (Davenport, 2014) and therefore only a short-term distraction (Dunnette, 1966). Thus, to provide some guidance in this domain, this chapter begins with an exploration of the concept of big data, including the introduction of a framework of big data functions based upon current applications in staffing. Next, we explore each of the dimensions of that framework by presenting case studies drawn from the experiences of I-O psychologists working with big data in the area of staffing, each case study paired with literature-based exploration of related cautions and new horizons. Finally, we draw conclusions regarding cross-functional benefits and risks.

A FRAMEWORK OF BIG DATA FUNCTIONS IN STAFFING

As with most new technologies with a significant interest from industry, there are many definitions of big data, although most share a common thread. In the *Harvard Business Review*, McAfee and Brynjolfsson (2012) describe the most common breakdown of big data, defining its three key features as *volume*, *velocity*, and *variety*. First, volume refers to the quantity of data analyzed,

which is sometimes expressed in exabytes. An exabyte is 1,000 petabytes, a petabyte is 1,000 terabytes, and a terabyte is a 1,000 gigabytes. Thus, a single exabyte of nothing but Microsoft Word documents would contain approximately 50 trillion of them. According to IBM (2015), 2.5 exabytes of data are created worldwide each day; tapping into this vast resource is part of what big data proponents seek to accomplish. Second, velocity refers to the speed of both data creation and analysis. In addition to the speed of data creation described above, using big data analytic techniques, real-time analysis of any phenomenon of interest might be observed. For example, an internal, employee-directed social network site might be automatically monitored for emotional content using real-time text mining. In doing so, management could get an up-to-the-minute estimate of the emotional state of their employees. Third, variety refers to the many different forms big data might take. Although text data such as electronic communications and electronic records are the most common, meta-data such as Internet history, radio-frequency identification (RFID) data such as physical location and the amount of time spent in various parts of an office, global positioning satellite (GPS) data, audio and video data, and other types of data are now often collected and analyzed together (Cumbley & Church, 2013).

Although this set of three characteristics is commonly found in big data definitions, additional dimensions are often added, and these dimensions and therefore definitions of big data vary by discipline (Hitzler & Janowicz, 2013). For example, *value* refers to the specific explanatory power of information to solve specific problems or challenges, *veracity* refers to the uncertainty surrounding information collected, and *variability* refers to the often inconsistent nature of collected data. Some authors have further defined big data as data that cannot be meaningfully processed or analyzed using conventional approaches (e.g., Dumbill, 2013), which includes all standard statistical software commonly applied in staffing, such as SPSS and SAS. From this perspective, big data is defined largely by the necessity of distributed processing, a technology involving tens, hundreds, or thousands of computers running in tandem, called a cluster, to achieve the high speed and accuracy of data handling necessary to meet whatever demand exists. For example, during a sales event in 2015, online retailer Amazon.com sold 398 items per second (Garcia, 2015), each purchase requiring a significant amount of data to be accessed and updated at a data processing rate currently impossible for a single personal computer to achieve. Even among data scientists, those academics and practitioners most directly connected to big data, the definition of big data—and for that matter, data science—is currently contested (Provost & Fawcett, 2013).

Given these disagreements, a precise and agreed-upon definition of big data may be less useful in the staffing context than a framework demonstrating how the various technologies typically involved in big data might be used to improve organizational functioning. Advantages to big data are proposed to be quite broad. For example, a report by the McKinsey Global Institute described five general organizational advantages to incorporating big data: (1) increased transparency and usability of data, (2) increased accuracy and detail of data, (3) increased specificity of data, (4) improved decision-making based upon data, and (5) improved research and development pipelines within organizations (Manyika et al., 2011). From the perspective of an I-O psychologist or other staffing specialist, many of these supposed advantages to big data likely seem quite familiar. The introduction of quantitative measurement to management formalized the data-gathering process, and data regarding human resources are now commonly collected and maintained in order to make the best decisions possible regarding organizational personnel. In this sense, organizations already collect transparent, usable, accurate, detailed, specific data about human resources that can be used to improve decision making in order to ultimately increase value. If big data is to provide new value to staffing, it must measurably improve one or more of these properties beyond what is currently possible with the existing I-O toolkit. Data must be *more* transparent, usable, accurate, detailed, and/or specific in such a way that an advantage is gained, despite increased costs due to specialized computer programming expertise and the use of complex computing systems.

To maximize the apparent value of big data in these ways, we have developed and present here a framework of big data functions based upon its four major application areas in staffing. These areas are not intended to be an exhaustive list of the ways in which big data might be used in staffing. We also do not mean to imply that these areas are orthogonal. Big data applications typically apply multiple technologies simultaneously. Instead, we have created this framework to illustrate the most common ways that big data technologies are currently applied in

order to highlight where industry staffing professionals believe the greatest added value might be achieved. Thus, we contend that if there is value to be found for staffing, it is likely to be in one of these areas. However, this does not preclude the creation of new application areas in the future, nor does it preclude additional uses beyond those we describe.

The first of these areas, big data gathering, refers to the use of big data technologies to collect data that was never before realistically collectable. One of the most relevant applications to staffing is the extraction of data from both external social media platforms, such as LinkedIn and Facebook, and internal social media platforms (Landers & Goldberg, 2014). Using social media, current employees and job applicants create lengthy and complex attitudinal and behavioral records that are often accessible to organizations. In the case of Facebook, Twitter, and other personal social media platforms, this behavioral record is quite focused upon the personal life of the person in question but still may contain job-relevant information. For example, using big data analytic techniques on a sample of 86,220 Facebook users, researchers developed an algorithm that can predict self-report personality ratings from Facebook likes better than judgments by their friends can (Youyou, Kosinski, & Stillwell, 2014), which the researchers framed as "computers outpacing humans in personality judgment" (p. 1036). Alternative measurement methods like this potentially bring many advantages to the measurement of predictors of performance, such as data collection speed and reduced fakeability, in comparison to self-report surveys.

The second of these areas, big data storage, refers to the use of big data technologies to maintain massive databases, which are far larger than any traditional staffing data sets. Most relevant to staffing is the incredible quantity of data now captured by wearable electronic devices, such as electronic employee badges. Wearables as a technology have existed for some time, although primarily for the purpose of personal healthcare (Lymberis, 2003), with only a recent expansion into broad consumer and enterprise applications, such as smartwatches. Wearables may be considered one part of a broader concept called the Internet of Things, which refers to the increasing movement toward providing Internet access to a wide variety of objects that have never before had Internet access (Xia, Yang, Wang, & Vinel, 2012), including household appliances. In the case of wearables at work, sometimes called enterprise wearables (Sacco, 2014), an employee badge might collect data on the specific location of the wearer throughout the workday, the doors accessed throughout the building, the other people with whom that person has been in close proximity, any sounds that resemble spoken words from which the speaker of those words can often be identified, and other such information. Once the badge passes within proximity of a reader, strategically located throughout the office, this information is uploaded to a central location. Although such information is now often collected in corporate environments where electronic badges are used, it is unclear what value this information might hold for the organization. Perhaps more importantly, the liability of holding onto this information is also unknown. This liability may be legal or may be felt more as a violation of employees' sense of privacy, trust, and respect in their relationship with their employer.

The third area, big data analytics, refers to the wide variety of data analysis techniques that have been developed as a result of the complexity of big data. Perhaps the most prominent of these techniques are machine learning and data mining (Chen, Chiang, & Storey, 2012). In contrast to I-O psychology's "theory-first" deductive approach, data scientists approach data holistically and inductively, seeking ways to simplify the data and extract meaning. Theory is the result of this process, not the cause. Where psychology relies on the deductive approach to minimize the degree to which conclusions are drawn based upon statistical artifacts, data science has developed statistical approaches to do this post hoc, generally based upon multivariate statistical approaches familiar with I-O psychologists. For example, in staffing, linear regression is often used to develop an equation predicting job performance from selection predictors. Used this way, regression works reasonably well with a relatively small set of predictors. In the case of big data, however, the number of potential predictors might increase to a few hundred or thousand. Because regression prioritizes explanatory power when adding predictors to a model, such an analysis would likely result in a high degree of capitalization on chance. To deal with this problem, data scientists might use a least absolute shrinkage and selection operator technique to be used in combination with linear regression in order to maximize prediction while also maintaining parsimony (Tibshirani, 1996). With this technique, all possible combinations of predictors can be modelled simultaneously to determine the tradeoff between explanatory

power and parsimony, allowing a data scientist to pick the regression model that best achieves a desirable balance. Models like these can also be developed automatically, programmatically, and iteratively, using a wide range of statistical techniques.

Many, although certainly not all, big data analytic techniques are distinct but recognizable cousins to statistical approaches common in I-O psychology. One commonly discussed technique in data science is machine learning, which is commonly used to sort ambiguous data into categories. I-O psychologists are generally familiar with two statistical techniques that accomplish the same general goal: factor analysis and cluster analysis. In both of these approaches, patterns within data are used to develop a broader classification scheme that can be used later for other purposes. For example, the Big Five personality traits were originally developed in part by using factor analysis to sort personality judgments into categories based upon words found in English that can be used to describe people. Machine learning is often employed to do similar sorts of categorization, but with a much greater degree of flexibility and autonomy. For example, the Big Five traits might be identifiable by programming a computer to comb the Internet (Landers, Brusso, Cavanaugh, & Collmus, in press), identifying words that appear to be descriptors of people based upon their position in each sentence. Next, the computer could iteratively process every sentence it identified to determine which personality words tend to cluster together, aided by a database of synonyms for reference, developing a personality model as it went. As the computer continued to collect more data, it would incrementally refine this model to better represent the data it has already collected, correcting for chance variation increasingly over time based upon the size of the data set at the time. In this way, such a machine could develop the Big Five automatically and algorithmically using cutting-edge technologies, yet this approach has the same conceptual basis as what was done by psychologists in the 1930s (i.e., Allport & Odbert, 1936).

The final area, big data visualization, refers to the use of interactive displays of data that allow viewers to parse the meaning of data in highly complex ways without any data science expertise. Data visualization was developed in part to help people make sense of fleeting data before their value disappears (Keim, Qu, & Ma, 2013). For example, in the time it might take for a data scientist to analyze data and develop a report to interpret its findings, the competitive advantage that might be gained for that organization could be lost. Additionally, in an environment where new data are created constantly and old data may become obsolete in a very short time, such a report may even provide faulty or harmful recommendations. Using data visualizations, key decision makers can explore summaries of data in real time, as those data change. Such a person could click-and-drag to explore organization-wide patterns to draw insights or "zoom in" to see differences between individual organizational units. In the context of enterprise wearables, a manager might be able to see the current locations of all employees in a real-time interactive map but also obtain real-time summaries of how many employees are at their desks, how many are at the water cooler, how many are in the restroom, and how many are on smoke breaks.

As demonstrated above, the possibilities of big data are far-reaching. However, reality often lags behind possibilities. In the next four sections, we will explore each of these four functions of big data—gathering, storage, analytics, and visualization—by presenting an anonymized case study describing how I-O psychologists working in staffing have utilized big data. After each of these case studies, we consider those applications from the perspective of available research literature within both the staffing literature and data science literature to identify strengths, weaknesses, and future directions.

BIG DATA GATHERING

Case Study

A moderately sized, regional organization grew dramatically by acquisition over a period of five years from 3,000 people into a geographically distributed, global organization of over 10,000. The original company had enjoyed a favorable reputation in its community as a good employer, and staffing processes had been fairly simple, based largely on employee referrals and a good relationship with the local university. Those close relationships meant that, in most cases, new

applicants had been known to the company as interns, scholarship recipients, or secondhand through recommendations from their professors, and selecting "the best" among them had seemed quite straightforward, given the work samples available from internship performance and classwork performance, which was available directly or vicariously.

The company's expansion had been based on product line complementarity, and none of the recruiting infrastructure of good employment brand and close university relationship was present in the newly acquired firms. In fact, in most cases, the existing goodwill that the legacy firms had enjoyed in their communities was damaged by the acquisition. Furthermore, as is common after acquisitions, there was a spike in attrition within most of the companies the organization acquired. Thus, the organization had to simultaneously address several challenges in its previously sleepy staffing function. They needed to understand who was leaving, why, and where they were going. They needed to understand their own employer brand and position in the landscape of employers, and they had to figure out how to recruit mid-career professionals for the first time.

A team of three talent analysts built a big data strategy to address these challenges, using multiple sources of data, including social media. In the first phase, they tackled the problem of attrition and talent flows. To do this, they applied natural language processing, a technique to extract meaning from text data, to exit interview notes and survey data, next applying machine learning to understand who was choosing to leave and what key drivers of attrition were. They then collected a large volume of social media data, primarily via LinkedIn's tools, to identify where their former employees had gone. Their review also revealed that a handful of employees had left after the acquisition but later returned. These people were asked to provide interviews.

The second phase of their work was understanding their employment brand in the marketplace. The team analyzed social media ratings and comments regarding their company, the companies that had absorbed most of their exiting employees and were thus their biggest talent competitors, and the legacy company names from prior to the acquisition. This provided insight on what at least a sample of employees and former employees viewed as important in their employment relationship and how each of the companies studied fared in the eyes of employees. This gave the researchers an idea not only of their competitiveness in the marketplace but also key assets they could highlight in their employment brand communications and key limitations they could work to address within the company. This information was especially helpful as they considered, for the first time, recruiting mid-career professionals. Here, the researchers reached out to recruiters from the acquired organizations for best practices and supplemented those practices with insights from the social media review.

From all of these efforts, the researchers learned that the company's generous leave policies, including unlimited vacation and periodic sabbaticals, were very highly valued by employees, especially by emerging professionals. However, employees, especially those mid-career to senior leaders who left, were frustrated by what they saw as very limited opportunities for influence and promotion in a company where interpersonal trust, based on many years of working closely together, was key to decision making. Based upon these findings, the company invested in highlighting its generous leave as a key employee benefit early in the recruiting process. The researchers also took their discovery around departing employee frustrations to company leadership and influenced organizational structure to visibly include a critical mass of leaders from outside the original, acquiring organization.

Finally, the company invested in a specialized leadership recruiting team that extensively used professional social media to identify candidates with appropriate skills and experience. The researchers built a playbook that highlighted the organization's employee value proposition in contrast to those of key talent competitors, and trained the leadership recruiters to subtly use that perspective in wooing candidates, highlighting key areas where the company was attractive as an employer compared to talent competitors. As they worked to improve their ability to identify, attract, hire, and retain these mid-career employees, they continually revisited their original analyses, periodically re-examining exit trends, talent transfer rates among the key companies with the highest talent flows among them, and social media sites. They adjusted and enhanced their employer branding materials on their company pages on professional social media, as well as in college recruiting in response to new information, and watched with pleasure as their employer ratings improved on social media sites. As their sophistication with social media grew, they also monitored visits to their company pages on professional social media and noted what changes to employment brand messages resulted in better candidate flows (Table 43.1).

TABLE 43.1

Summary of "Big Data Gathering" Case Study

Staffing Application	The organization needed to gather information about the causes of employee turnover.
Limitations of Small Data	Exit interviews are time-consuming and resource intensive, relying upon thoughtful answers in a face-to-face setting, which does not promote frank honesty.
Advantages to Big Data	The harvesting of social media data, in combination with machine learning and natural language processing, allows organizations to develop insights about turnover based upon not-previously-accessible information. Employee in-flow and out-flow analysis based upon this data helps draw conclusions regarding motivation.
Cautions	The ubiquity of social media does not necessarily solve more fundamental sampling challenges. Furthermore, very large samples are required for predictive accuracy.

Conclusions, Cautions, and New Horizons

As shown in this case study, big data gathering techniques can be used to collect multiple dissimilar types of information, such as text extracted from interviews and social media streams, to produce a single model from which insights can be drawn and predictions made. Such data collection is particularly useful in this context for two reasons. First, the collection of unstructured data from social media enables follow-up from ex-employees whose opinions would normally be inaccessible to the organization. Second, because there is relatively little theoretical guidance on what specific human resources policy changes might be perceived as problematic after a merger, this approach enables high-quality, data-driven decision making. The results from this approach will be highly organization-specific, but a highly organization-specific solution is precisely what was needed to solve this problem.

Importantly, big data techniques do not avoid the traditional challenges of sampling. As Harford (2014) notes, it is seductive to think of big data as "N = All" yet this is a risky assumption. Landers and Behrend (2015) describe the considerations associated with using convenient sources of data like these, big or small. Importantly, relationships of interest must not covary with membership status in the convenient sample, or results from that sample will be biased. In this case, it would be important to ensure that the reasons shared on social media were common among all employees who left the organization and not unique to those complaining on social media. In this case, the organization saw improvements in their staffing function, but the benefits might have been even greater with a better source of information—perhaps even one from small data, if such data had been otherwise attainable.

For big data gathering of this type to be effective, the data source must also be quite large. Thus, the organization also benefited from its own size, which enabled a significant amount of social media data to be collected. In an organization with a low absolute turnover rate, big data of this type may be less useful since fewer data are likely to be available. Much as with I-O psychology's mainstream selection techniques, small samples and small employee populations add a great deal of noise to available data, decreasing the evident value of many staffing practices (Sackett & Arvey, 1993). Small organizations may find greater value in gathering big data from public but highly relevant sources, such as those that can be geographically targeted. However, this introduces generalization challenges.

Specifically, many of the scaling challenges associated with synthetic validation apply similarly to big data. Synthetic validation refers to validity evidence gathered by logical inference to draw conclusions about particular jobs based upon broader, non-organization-specific validation efforts when a traditional concurrent or predictive validation study is not feasible due to either small sample size or lack of criterion data (Scherbaum, 2005). Similarly, big data of a desirable type may not be available from current employees. In such cases, staffing specialists will need to determine how dissimilar the data can be yet still provide useful information. In this case study, the organization decided that whatever information was posted on social media by current employees of competitors and its own

ex-employees was trustworthy. The only statistical test that could determine if this assumption was valid would not be necessary if the population data necessary to conduct it were available, so this will always be an assumption for practitioners to make. It is one that should be made cautiously.

The approach taken here also highlights another risk of big data. When researchers assume that data created in the past must contain all the answers needed in the future, those biases become part of the conclusions drawn. Specifically, big data is typically previously collected data. Its availability may discourage researchers from considering creative, alternative solutions that are not present. In this case study, the talent analyst team started with an assumption that exit interviews and social media would highlight the most efficacious solutions for the organization. Just as when using traditional research strategies, the design of the study that created the data set drives the conclusions that can be drawn from it. The original data collection decisions that created big data, such as the social media case described here, are rarely under the researcher's control, which introduces a degree of risk. Given this, we recommend researchers considering big data approaches to their talent problems carefully consider what creative solutions not relying on existing data might be employed. Ideally, a combination of both forward- and backward-looking data should be used as the basis for decision making.

BIG DATA STORAGE

Case Study

At a large organization, the staffing group was having a problem making good hires for sophisticated manufacturing technician roles that required specific manual skills and high degrees of both teamwork and coordination. About 30% of new hires did not successfully complete their 90-day evaluation period and were terminated before completing it. Although the numbers of employees in these roles were not large, errors were costly, and it was beneficial to the organization to limit the risk of poor performance even during this 90-day trial period. To improve the number of successful applicants, staffing decided to capture actual performance of job tasks and test this performance in a simulation to be made a key part of their selection process. To do this, the organization implemented wearables, which were intended to collect a massive volume of information about performance.

The project began by identifying a core group of successful employees. These employees volunteered to spend 40 total hours over three months working with the project team in order to build a realistic simulation of the essential functions of the job. The volunteers wore a wrist-mounted device on their dominant hand that measured specific locations and proximity to equipment and interactions with that equipment, as well as interactions with team members. The volunteers also wore a head-mounted, eyeglass-style device that tracked eye movements and thereby measured attention to specific pieces of information. The level of detail enabled by the wearable device vastly increased the number of available variables for measurement and prediction. Both devices were lightweight and judged to be non-intrusive. Personal biometrics, such as stress responses, were not measured, out of a concern that employees and job candidates might perceive it as a violation of their privacy.

Once the simulation was built and the quality of the measurement system had been well established, additional content-related validity evidence and also concurrent criterion-related validity evidence were gathered by asking additional employee volunteers to spend one hour participating in the simulation, wearing both the wrist-mounted and head-mounted devices. These employees were then asked how accurate and relevant the experiences in the simulation were, and data collected from their wearable devices were compared to metrics of actual on-the-job performance. The predictive model developed during the first phrase was refined based on this larger set of results.

The first group of candidates to complete the simulation was a test group, used as part of a predictive validation study. For this group, the simulation was used as a realistic job preview, but the results were not shared with interviewers and thus were not considered in the final hiring decisions. This way, data from applicants could also be used to refine the predictive algorithm. After a few final adjustments were made to the predictive model, it was incorporated into the hiring process.

TABLE 43.2
Summary of "Big Data Storage" Case Study

Area	*Summary*
Staffing Application	The organization wanted to understand job performance at a high level of detail to better predict those behaviors.
Limitations of Small Data	The sheer quantity of tiny, difficult-to-observe pieces of information makes it difficult to know a priori which of them are actually relevant to job performance and in what combination. Even if specific information could be chosen, problems with rater training and rater accuracy are significant.
Advantages to Big Data	The ability to store a massive amount of data enables a model to be built based upon that massive amount of data. The specific challenges associated with identifying what is relevant because to a machine learning algorithm, all of the data can be considered simultaneously.
Cautions	Privacy is a concern when storing massive data because many (perhaps most) people are not aware of how much data is really collected. Organizations may incur heightened legal risk if opposing counsel subpoenas those data and mines for chance relationships. Data security is also a major concern and requires significant technical expertise.

During the first year of implementation of the wearable-enhanced simulation, the failure rate during the evaluation period was reduced by over 60%; that is, the failure rate during the evaluation period went from 30% of new hires to 12% of new hires. Furthermore, the cost reduction associated with reducing errors by new hires paid for the program investment within the first 10 months of program implementation (Table 43.2).

Conclusions, Cautions, and New Horizons

As illustrated in this case study, big data techniques can be used effectively as additions to existing selection and training techniques already well-known in staffing. In this case, big data storage enabled the collection of a wide variety of data types at a high velocity in a simulation, itself a method already commonly used for both selection and training when high-fidelity representation of job tasks is a priority (Boyce, Corbet, & Adler, 2013). Importantly, the addition of big data does not diminish the importance of a traditional and comprehensive validation process. Here, content-related validity evidence was collected from both an initial pool of subject matter experts and later from a broader employee sample. Criterion-related validity evidence was also collected, first in a concurrent design and later in a predictive design, as commonly recommended by selection experts (Society for Industrial and Organizational Psychology [SIOP], 2003). The inclusion of wearables does not change this; it only adds greater breadth and depth to the type of data collected.

Although wearables as used in this study increased both the breadth and depth of data collected from those participating in the simulation, such data are not necessarily useful. If data relevant to the problem to be solved are never collected and stored, no degree of analytic complexity will be able to extract useful information from them. Thus, it is important to consider precisely what kind of data is being stored by the devices creating those data. In this case, the wrist-mounted devices worn by participants primarily captured distances. These distances were calculated based upon the locations of other wrist-mounted devices and stationary objects broadcasting their location. If distances were not relevant to job success, then the distance data stored by the wearables would be effectively useless, despite the vast size and complexity of those data. To prevent the collection of low-value data, it is therefore recommended to carefully link existing theory and research to each particular problem to be solved. With big data, size alone is insufficient.

Inspired by this case, we identified three other major cautions related to the long-term storage of vast quantities of data. First, privacy is a major concern. Existing research in selection already

notes the impact of perceived privacy violations on applicant reactions (Bauer et al., 2006), and such violations are much easier to make when a firm's big data philosophy involves the collection of as much and as varied data as possible. Importantly, perceptions of privacy violation and actual privacy violations are distinct. Applicants may perceive that their privacy has been violated when it in fact has not and vice versa. Big data that are collected surreptitiously will not influence applicant perceptions until the collection effort becomes known; however, such a policy creates the potential for a highly publicized public outcry when it is discovered (e.g., Hackett, 2015). Even in the relatively low-risk case study described here, in which big data were only collected on job incumbents and used to generalize to applicants, staffing specialists were concerned that the wearables might collect information that their employees would see as "off-limits." Such potential privacy violations should be carefully considered when any organization plans to create big data, and the targets of planned big data gathering efforts should be consulted before any databases are actually created.

Second, there may be a degree of legal risk associated with the collection and maintenance of vast quantities of big data. For example, if the staffing specialists in this case study had provided wearables to all incumbent employees, rather than just targeted individuals during the simulation development process, a vast database containing all movements of all employees over an indeterminate amount of time would have been created. In certain types of legal challenges, organizations might be required by subpoena to provide their big data to opposing counsel, as is common in adverse impact cases (Guion, 2011). Because big data are so complex, there are many ways to analyze them without generally agreed-upon standards, making competing interpretations likely (Bollier, 2010). For this reason, we recommend organizations only collect those big data that are needed for specific purposes, and retain them only as long as necessary for those purposes, echoing older recommendations regarding small data (Binning and Barrett, 1989).

Third, precautions should be taken to ensure that big data are stored securely. Small data sets are generally easy to anonymize (Ghinita, Karras, Kalnis, & Mamoulis, 2007), limiting the damage that can be done if those data sets are accessed by unauthorized personnel. Even in cases where data are somewhat more complex, such as personnel records, there are many well-established security practices to keep those data safe. In contrast, the scope of big data means that information may be stored across many systems, many user accounts, many physical locations, and potentially many organizations. Each of these is a potential security breach point and must be treated with the same care as any other single data source, also taking care to meet the requirements of the various legal systems within which those data exist. This is less of a concern for large organizations that are already accustomed to maintaining large, secure data warehouses. In these organizations, the storage of big data requires only an expansion of existing resources. In smaller organizations without existing standards-compliant secure data storage, a great deal of caution must be exercised to ensure that security standards are met as data storage capacity is increased to handle these new requirements. For such organizations, we instead recommend cloud-based solutions such that all potentially sensitive big data are stored and secured by organizations specializing in data warehousing and data security. Importantly, such a strategy is still not risk-free. Whereas cloud storage is likely to have superior countermeasures and protection, it is also a much more tempting target to hackers than a lone organization's databases.

BIG DATA ANALYTICS

Case Study

A global employer became concerned that its keyword-search-and-filter-based process for identifying job candidates within its Applicant Tracking System (ATS) was missing successful candidates. The company received hundreds of thousands of applicants per year across several job titles and ultimately hired more than 1,000 employees each year. These parameters led them to believe that artificial intelligence could add both efficiency and accuracy into their candidate identification process. In this case, the type of artificial intelligence targeted was machine

TABLE 43.3
Summary of "Big Data Analytics" Case Study

Area	*Summary*
Staffing Application	The organization wanted to improve its recruitment pipeline to identify and target higher-quality applicants at a faster rate.
Limitations of Small Data	Minor indicators of success often go unnoticed by recruiters, who are also influenced by a variety of personal biases that may influence their judgments. Recruiters also cannot respond to shifts in the labor market without interpreting a significant amount of business intelligence.
Advantages to Big Data	Algorithms can identify and make judgments about candidates automatically without intervention, responding to labor market shifts as they occur. Algorithms can also respond to internal personnel records as they change, resulting in the most accurate predictive model at all times.
Cautions	Although these models are powerful, it is important to maintain existing well-supported I-O processes. Big data recommendations should be validated and treated as a distinct hurdle in the selection process. Feeding data to an algorithm will result in predictions based upon those data, so care is needed when considering what sort of data to feed.

learning, a process by which algorithms are developed iteratively and automatically to produce a predictive model (Kotsiantis, Zaharakis, & Pintelas, 2007).

Given high direct and replacement costs of attrition at this employer, the company considered both hiring rates and retention rates when identifying five particular job titles for a pilot project. Within each of these job titles, a group of employees was identified as "successful" based upon two characteristics: (1) a tenure of at least two years and (2) current high job performance records. The original job applications were used to train a machine learning model tasked with identifying this group. Inputs for the model included both resume data and process data, such as the channel by which the person applied (e.g., as an employee referral, a participant in a job fair, a student at a target school). The model was then refined by providing data on candidates who were hired but not in the successful group. These were candidates with poor performance and those who left voluntarily. This helped develop a set of markers for candidates at risk of being false positives. Separate models were built for each of the target positions.

Given the volume of candidates present in the ATS, it was assumed that, in addition to the false positives identified in the step above, the ATS contained a number of false negatives. To investigate this, the machine learning algorithms already developed were next applied to candidates who had not been hired but remained in the ATS. Specifically, this assumed that the previous approach was overlooking good candidates who were already in the applicant pool. The machine learning approach was successful: the algorithms were able to identify additional candidates who were likely to perform well but who had been overlooked initially. These applicants were then hired and did generally perform well. This information was then used to further improve the algorithms.

The organization made a choice to use the algorithms as a complement to its existing, recruiter-driven process, rather than rely exclusively on the machine learning approach. This was primarily to ensure that the organization remained nimble as the industry and competitors evolved. The organization was concerned that exclusively relying on a backwards-looking approach would cause it to miss market shifts. Thus, application of the machine learning algorithm was used as a final step in preparing candidate slates, rather than the first one (Table 43.3).

Conclusions, Cautions, and New Horizons

As illustrated in this case study, big data analytics can be used to improve the prediction of existing employee selection processes. Big data approaches do not need to replace existing practices and can be used as a supplemental selection tool. What remains unclear are the consequences

of this improved prediction. I-O psychologists go to great lengths to ensure a high degree of construct validity for the measures they choose (Binning and Barrett, 1989). This is done, first and foremost, to ensure that prediction of job performance is based upon a well-defined characteristic of each applicant. If a conscientiousness measure is used, we must be confident that the measure is in fact one of conscientiousness. This value is reflected in all commonly accepted measurement guidelines (e.g., American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014). Practically speaking, this is in part to reduce the risk of loss in litigation; in the event that a selection system is legally challenged, a clear record of validation efforts is necessary to defend it (Guion, 2011).

The consequences of ignoring the validation process and instead entirely relying upon a machine learning algorithm, the internal details of which are often unknown to their users, may be significant. If any variable contained within the data is correlated with group membership in a protected class, that variable will be included in any resulting algorithms and result in biased selection. For example, if information found within a "Personal Interests" section of a resume provides useful information in the prediction of job performance, but the presence of such a section is by chance correlated with sex, a sex bias will be introduced into the resulting algorithm. Selecting on anything highly correlated with membership in any protected class will result in significant legal risk in the United States (Hough, Oswald, & Ployhart, 2001), which leads us to conclude that machine learning algorithms cannot be used indiscriminately in selection systems. To minimize potential problems, we recommend that organizations only use machine learning algorithms in hiring as a distinct hurdle, as described in the case study. This way, the results of recommendations from the algorithm can be validated independently, as is recommended in hurdle systems (Mendoza, Bard, Mumford, & Ang, 2004). If a problematic bias is discovered, the algorithm can be modified and the effects observed directly.

This problem is reminiscent of the days of so-called dust-bowl empiricism in I-O psychology, an era when any characteristic of a person that improved prediction of job performance was considered a reasonable hiring tool (Bryan & Vinchur, 2013). Many of the problems associated with that approach have reappeared here. In particular, because big data invites the inclusion of any and all even vaguely relevant data sources to improve its algorithms, job relevance of included data may be quite low. In the case above, process data were restricted to sources that the staffing team believed likely to aid in prediction, such as source of referral. However, a much broader array of process data could be collected and included in the algorithm, including the amount of time spent on individual web pages in the application process, the font size used on the resume, or any other such discrete piece of information provided by the job candidate. Anything given as input to the machine learning process to improve its algorithm's prediction may be used. Although it may in fact improve prediction, the lack of job-relatedness may be both legally and ethically problematic. To avoid this, we recommend only providing input to the machine learning process that is theoretically consistent with the prediction of job performance. In the case above, referral source was included, which has a supporting research literature (Zottoli & Wanous, 2000). Specific times spent on application pages were not.

Machine learning is closely related to another concept called data mining, which brings somewhat different challenges. In contrast to the traditional descriptive and inferential statistical approaches commonly used in staffing, data mining is a more flexible, computationally driven approach to understanding data (Hand, 1999). In a data mining approach, algorithms are developed by a researcher to identify patterns in data and build predictive models; automation might be used but is not necessary (Olson & Delen, 2008). Machine learning identifies such patterns and builds upon them automatically; in short, the researcher creates the intelligence, and the intelligence creates the algorithm. Data mining brings many of the same advantages and disadvantages of machine learning described above; however, the more hands-on role of the researcher potentially mitigates some of the disadvantages. Staffing specialists with knowledge of both data mining techniques and I-O psychology practices may be able to blend the best of both approaches, although this has not yet been demonstrated in the research literature. Most papers in this area to date have been written by data mining researchers (e.g., Chien & Chen, 2008; Cho & Ngai, 2003).

Big data analytic techniques are evolving at a rapid rate. The community tends to be practice-oriented, so new research is not always published in traditional outlets. Additionally, as is common in many fields related to and including computer science, the primary outlet for new research by those developing these techniques tends to be academic conferences. As a result, staffing specialists who are more familiar with traditional statistical approaches are likely to have difficulties both accessing and judging the quality of new research in big data analytics. Although some efforts have begun to appear related to big data research in staffing, the literature is quite sparse in comparison to the literature in data science broadly. As a result, for those seeking to implement big data analytics, we currently recommend seeking out and collaborating closely with professional data scientists who specialize in this domain, although this may change over the next few years as resources more accessible to staffing specialists are developed.

BIG DATA PRESENTATION AND VISUALIZATION

Case Study

A complex global organization with hundreds of standard job titles and dozens of major locations in multiple countries wanted to improve their overall staffing processes, including both recruitment and selection. Due to the complexity and volume of data, the organization turned to data visualization in two projects to help identify important patterns and to enable dynamic exploration of the data by organizational stakeholders without significant statistical or analytics expertise. In taking this approach, the researchers hoped to empower decision makers to act on data without the traditional complexities of statistical reporting.

Their first project was intended to improve staffing processes. In this case, the organization built a visualization that displayed key process steps, recruiting channels, job titles nested into job families, geographies, levels, and recruiter caseload for each job requisition. The initial data display showed the global average time and standard deviation of time for each step in the recruiting process. Users could then click on each process step to drill down to any combination of variables of interest. This enabled users to quickly identify outliers, as well as best and worst in class, within each class and for each set of variables being targeted. The organization was able to explore thousands of combinations and visually identify three process steps that introduced the greatest variability. The best-in-class examples were then used as prototypes to build new standard processes.

The second project was intended to better understand the current workforce and available labor markets in order to build new recruiting strategies. For this visualization, a map view of the organization was created showing unit populations and recruiting trends within each population. After consideration of the most challenging areas from this visualization, additional recruiting times and barriers data were added to better understand which strategies would be most effective in these challenging areas. Next, external labor market data, using census data and other sources, were added to enable the organization to identify which positions could be best served with local searches and which should be bundled together and addressed with a multisite, national or global search. This approach maximized efficiency in search times and cost in terms of relocation and retention. Finally, the organization analyzed efficiency for each of the recruiting channels and strategies at the local and national level in order to identify optimal criteria for each recruiting strategy. Specifically, the organization was able to explore which strategies best served each combination of recruiting circumstances.

In doing so, the organization built a recruiting strategy around insights gleaned from visualized data. This increased the degree of data-driven decision making in the organization, because before this point, the personal insights and creativity of executives and recruiters typically drove recruiting strategy. Not much attention was generally paid to the key roles and groups of roles that were particularly hard to fill because the difficulty filling these roles only became obvious with the visualization. Based upon conclusions drawn from the visualization, the organization also established a satellite team near a particular university to capitalize on the flow of candidates from that school in that specific area. They furthermore segmented the recruiter organization; part of the team focused on efficiency and transactions in the areas with a highly liquid talent market, and the remainder focused on proactive, passive candidate recruitment in areas that were more difficult to fill (Table 43.4).

TABLE 43.4

Summary of "Big Data Presentation and Visualization" Case Study

Area	*Summary*
Staffing Application	The organization had such a large quantity of data that it was difficult to understand all aspects of the recruitment and selection pipeline simultaneously.
Limitations of Small Data	Traditional visualization and presentation of data involves taking a snapshot of current data relationships. These visualizations may become outdated quickly. Creating such visualizations is also usually the task of a data analyst, which adds a step between the collection of data and action based upon those data.
Advantages to Big Data	Big data visualization techniques enable data to be visualized live as changes occur. Instead of considering a snapshot of data now, a stream of data is considered as it is created.
Cautions	Many of the same downsides to small data visualization still exist with big data visualization. A great deal of power is provided to the visualization designer to dictate what viewers see and consider when making decisions. Unique to big data is the sheer quantity and variety of data, which exacerbates this problem. High quality design is critical.

Conclusions, Cautions, and New Horizons

As demonstrated in this case study, big data visualizations can serve as powerful analytic tools (Frankel & Reid, 2008). This is in stark contrast to the use of visualizations as a supplement to statistical analyses, where visualizations are unfortunately often an afterthought (Gelman, Pasarica, & Dodhia, 2002). Visualizations in both small and big data contexts can provide intuitive displays of complex data, enabling new insights if designed well. In the big data context, visualizations go beyond the capabilities of traditional figures and charts by adding interactivity. Those viewing big data visualizations can in effect create and interpret cross-sectional analyses at any level of specificity without ever looking at a number; thousands of static figures may be contained within a single visualization, and a person interested in one of those thousands of figures can view that one desired figure immediately and automatically upon request. Big data visualization tools can even be used with small data, although the added complexity is only worthwhile when this sort of interactivity would be valuable to the target audience.

The implication of this interactivity is that the specificity of insights is much greater, and this brings both unique opportunities and unique challenges. Because users may drill down to any of thousands of figures, and because the people creating visualizations rarely look at all possible permutations of figure enabled by those visualizations, drilldowns containing spurious results are likely. In the circle packing visualization found in Figure 43.1, for example, circle sizes represent the total number of employees in a large organization within each first-order job grouping (division), divided further based upon a second-order job grouping (product team). A user might click on any given circle to gain more specific information about that grouping and its subgroups, and then click within subgroups to get information about even smaller subgroups, as shown on the right side of Figure 43.1. In such cases, chance variation alone may cause a particular requested figure to misrepresent larger trends, a common problem with multilevel data (Klein, Dansereau, & Hall, 1994). In the same way that simple statistical tests can be misleading when contextual assumptions are not met, visualizations can be misinterpreted when viewers forget, ignore, or do not have access to the bigger picture. Because images in general are more persuasive than other more numerically oriented forms of information (Latour, 1990), visualizations have a great deal of power to misinform as well as inform.

Even when decision makers are prepared to consider visualization data from multiple perspectives to avoid this problem, the sheer quantity of information produced may be overwhelming. When a thousand different cross-sectional figures can be obtained, it is often unclear which should be prioritized and trusted. Humans are only readily able to consider a relatively small number of sources of information simultaneously in decision making (Payne, 1976); thus, the availability of so many figures may in this way be harmful. Statistical approaches were developed,

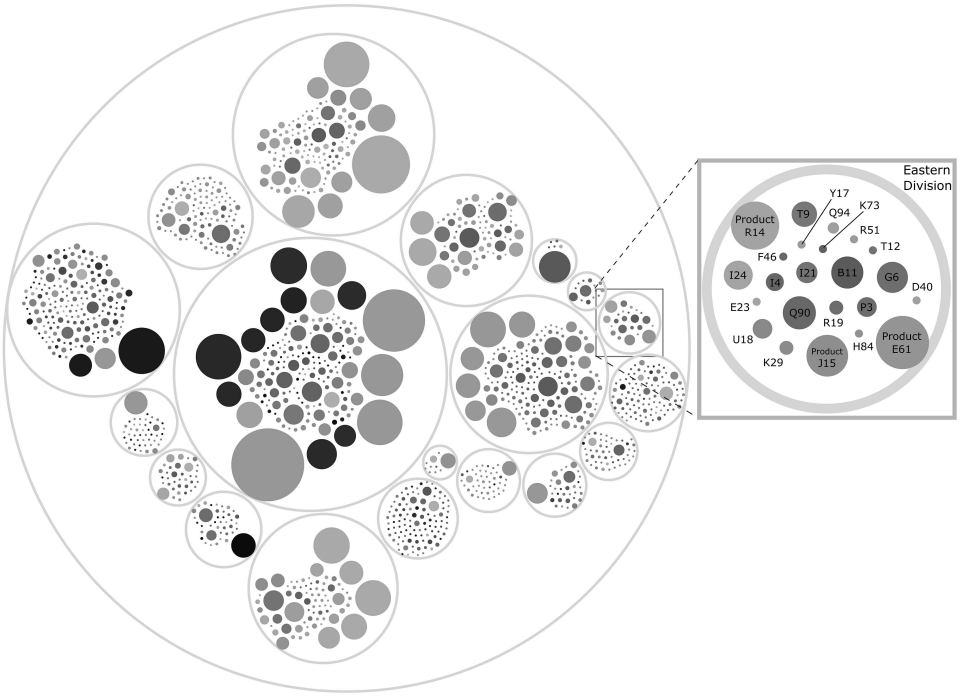

FIGURE 43.1 Sample Big Data Visualization

Source: Courtesy of Evan Sinar (Development Dimensions International).

in part, to simplify decision making from vast quantities of data. Although big data visualization tools may make it somewhat easier to sift through large amounts of data meaningfully, there is still a limit to human information processing.

Given these challenges, we recommend visualizations be used only in contexts where the specific affordances of data interactivity would aid in decision making. In such cases, the visualization should still be carefully designed to provide only relevant and actionable data to the viewer. Although excess variables can be easily included in visualizations, simplicity is still a virtue. Only those visualization options that are theoretically linked to the problem to be solved should be included. Because of the potential for misleading results, we also recommend that big data visualizations, when used analytically, only be used as the first step in the decision-making process, to then be followed up with small data investigations using traditional research methods.

CONCLUSION

In summary, we have presented four case studies highlighting each of the four functional areas of big data in staffing: gathering, storage, analytics, and presentation/visualization. Across these areas, there is a great deal of potential for staffing to be transformed by big data. We can now collect information we could never collect before at a scale we could never before collect it, applying a wide variety of analytic techniques based upon artificial intelligence research to identify patterns that can be acted upon. We can create interactive visualizations so that people with no statistical expertise can interactively and powerfully explore data, to make data-driven decision making well within the reach of even the most numbers-phobic organizational leader. This provides an incredible opportunity to increase the accuracy of both staffing decisions and staffing research.

There is also a great deal of potential to mislead ourselves. These techniques are quite powerful, bringing many opportunities to head down a harmful path based upon seemingly minor

decisions. The ease of data gathering means that far more data can be collected than are useful, encoding information with unclear value and potential legal risk. Big data storage is so inexpensive and vast that massive amounts of data can be stored essentially indefinitely. This can create a tempting target for hackers, yet sensitive electronic information cannot be stolen if it is not accessible to the Internet (or to big data practitioners). Big data analytics offer the ability to extract insights from data that were never before extractable, identifying subtle patterns of numbers that a human analyst running traditional analyses would likely never find, but these approaches are often quite brute force, extracting patterns in samples when no such patterns may exist in the population. Big data visualizations that enable non-statisticians to dive deeply into data also may create a false sense of security, and the type of information conveyed by such visualizations is entirely under the control of the visualization designer, who will likely make hundreds or thousands of small decisions along the path from raw data to a particular visualization.

Given this combination of potential and caution, we contend that the greatest value will be found at the intersection points between big data and traditional staffing research. When these two families of techniques are used in concert, when insights are discovered with big data and verified with the collection of in-depth small data, we can be maximally confident that the right decisions are being made. Echoing recommendations for mixed-methods research (Creswell & Clark, 2011), we contend that the convergence of multiple methods on the same recommendation is the best evidence to initiate a particular organizational intervention. When these multiple methods do not converge, it is time for further investigation; conclusions drawn from big data are neither inherently better nor worse than those drawn from small data. Instead, an interdisciplinary perspective will provide the answers organizations seek, and I-O psychologists, staffing specialists, and big data practitioners should try to build this perspective.

NOTE

1. We would like to thank Evan Sinar for his gracious contribution of Figure 43.1.

REFERENCES

Allport, G. W., & Odbert, H. S. (1936). Trait names: A psycholexical study. *Psychological Monographs, 47*(1).

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Bauer, T. N., Truxillo, D. M., Tucker, J. S., Weathers, V., Bertolino, M., Erdogan, B., & Campion, M. A. (2006). Selection in the information age: The impact of privacy concerns and computer experience on applicant reactions. *Journal of Management, 32*, 601–621. doi: 10.1177/0149206306289829

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478–494.

Bollier, D. (2010). *The promise and peril of big data* (Aspen Institute Report). Retrieved from http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

Boyce, A. S., Corbet, C. E., & Adler, S. (2013). Simulations in the selection context: Considerations, challenges, and opportunities. In M. Fetzer & K. Tuzinski (Eds.), *Simulations for personnel selection* (pp. 17–41). New York, NY: Springer.

Bryan, L. K., & Vinchur, A. J. (2013). Industrial-organizational psychology. In D. K. Freedheim & I. B. Weiner (Eds.), *Handbook of psychology, Vol. 1: History of psychology* (2nd ed., pp. 407–428). Hoboken, NJ: John Wiley & Sons Inc.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*, 1165–1188.

Chien, C-F., & Chen, L-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications, 34*, 280–290.

Cho, V., & Ngai, E. W. T. (2003). Data mining for selection of insurance sales agents. *Expert Systems, 20*, 123–132.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE Publications.

Cumbley, R., & Church, P. (2013). Is "big data" creepy? *Computer Law & Security Review, 29*, 601–609.

Dumbill, E. (2013). Making sense of big data. *Big Data, 1*(1), 1–2.

Davenport, T. (2014). *Big data at work: Dispelling the myths, uncovering the opportunities*. Cambridge, MA: Harvard Business Review Press.

Dunnette, M. D. (1966). Fads, fashions, and folderol in psychology. *American Psychologist, 21*, 343–352.

Frankel, F., & Reid, R. (2008). Big data: Distilling meaning from data. *Nature, 455*, 30.

Garcia, A. (July 16 2015). Amazon 'Prime' Day' shattered global sales records. *CNN Money*. Retrieved from http://money.cnn.com/2015/07/15/news/amazon-walmart-sales/

Gelman, A., Pasarica, C., & Dodhia, R. (2002). Let's practice what we preach. *The American Statistician, 56*, 121–130. doi: 10.1198/000313002317572790

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2007). *Fast data anonymization with low information loss*. Paper presented at the Proceedings of the 33rd international conference on Very large data bases, Vienna, Austria.

Guion, R. M. (2011). The legal context for personnel decisions. In R. M. Guion (Ed.), *Assessment, measurement, and prediction for personnel decisions* (pp. 163–207). New York, NY: Taylor & Francis.

Hackett, R. (June 9 2015) Massive federal data breach affects %7 of Americans. *Time Magazine*. Retrieved from http://time.com/3952071/opm-data-breach-federal-employees/

Hand, D. J. (1999). Statistics and data mining: Intersecting disciplines. *SIGKDD Explorations, 1*(1), 16–19.

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19.

Hitzler, P., & Janowicz, K. (2013). Linked data, big data, and the 4th paradigm. *Semantic Web, 4*, 233–235.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194. doi: 10.1111/1468–2389.00171

IBM. (2015). *Bringing big data to the enterprise*. Retrieved from https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Keim, D., Qu, H., & Ma, K-L. (2013). Big-data visualization. *IEEE Computer Graphics and Applications, 33*, 50–51.

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19*, 195–229.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2007). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*, 159–190. doi: 10.1007/s10462–007–9052–3

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizations, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology, 8*, 142–164.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (In press). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*.

Landers, R. N., & Goldberg, A. S. (2014). Online social media in the workplace: A conversation with employees. In M. D. Coovert & L. F. Thompson (Eds.), *Psychology of workplace technology* (pp. 284–306). New York, NY: Routledge Academic.

Latour, B. (1990). Drawing things together. In M. Lynch & S. Woolgar (Eds.), *Representation in scientific practice* (pp. 19–68). Cambridge, MA: MIT Press.

Lymberis, A. (2003). Smart wearables for remote health monitoring, from prevention to rehabilitation: Current R&D, future challenges. In R. Summers (Chair) & E. Carson (Co-Chair), *4th International IEEE EMBS special topic conference on information technology applications in biomedicine 2003* (pp. 272–275). Piscataway, NJ: IEEE.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 61–67.

Manyika, J., Chui, M., Brown, B., Buhin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition and productivity*. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Mendoza, J. L., Bard, D. E., Mumford, M. D., & Ang, S. C. (2004). Criterion-related validity in multiple-hurdle designs: Estimation and bias. *Organizational Research Methods, 7*, 418–441. doi: 10.1177/1094428104268752

Newell, A., & Simon, H. A. (1976). Coputer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Berlin, Germany: Springer-Verlag.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, *16*, 366–387.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*, 51–59.

Sacco, A. (Spring 2014). Enterprise wearables present tech challenges and management pitfalls. *CIO*, 44–45.

Sackett, P. R., & Arvey, R. D. (1993). Selection in small N settings. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 418–447). San Francisco: Jossey-Bass.

Scherbaum, C. A. (2005). Synthetic validity: Past, present, and future. *Personnel Psychology, 58*, 481–515.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Society for Industrial and Organizational Psychology.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Ur, B., Leon, P. G., Cranor, L. F., Shay, R., & Wang, Y. (2012). Smart, useful, scary, creepy: Perceptions of online behavioral advertising. In L. F. Cranor (Chair), *Proceedings of the Eighth Symposium on Usable Privacy and Security* (pp. 1–15). New York, NY: ACM Press.

Youyou, W., Kosinski, M., & Stillwell, D. (2014). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academies of Science, 112*, 1036–1040.

Xia, F., Yang, L. T., Wang, L., & Vinel, A. (2012). Internet of things. *International Journal of Communication Systems, 25*, 1101–1102.

Zottoli, M. A., & Wanous, J. P. (2000). Recruitment source research: Current status and future directions. *Human Resource Management Review, 10*, 353–382. doi: 10.1016/s1053–4822(00)00032–2