Smith, J. B. (1994). *Collective intelligence in computer-based collaboration*. Hillsdale, NJ: Erlbaum.

Steiner, I. D. (1972). *Group process and productivity*. New York, NY: Academic Press.

Sullivan, S. D., Lungeanu, A., DeChurch, L. A., & Contractor, N. S. (2015). Space, time, and the development of shared leadership networks in multiteam systems. *Network Science, 3*(01), 124–155.

Turek, P., Wierzbicki, A., Nielek, R., Hupa, A., & Datta, A. (2010). Learning about the quality of teamwork from Wikiteams. *Social Computing (SocialCom), 2010 IEEE Second International Conference*, 17–24. Retrieved from http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5590331

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequences of the human genome. *Science, 16*(291), 1304–1351.

Wilkinson, D. M. (2008). Strong regularities in online peer production. *Proceedings of the 9th ACM Conference on Electronic Commerce*, 302–309.

I-Os in the Vanguard of Big Data Analytics and Privacy

Adam J. Ducey
IBM, Hazelwood, Missouri

Nigel Guenole
IBM Smarter Workforce Institute, London, United Kingdom

Sara P. Weiner
IBM Smarter Workforce, Tucson, Arizona

Hailey A. Herleman
IBM Smarter Workforce, Frisco, Texas

Robert E. Gibby
IBM, Cincinnati, Ohio

Tanya Delany
IBM, Milan, Italy

In this response to Guzzo, Fink, King, Tonidandel, and Landis (2015), we suggest industrial–organizational (I-O) psychologists join business analysts, data scientists, statisticians, mathematicians, and economists in creating the vanguard of expertise as we acclimate to the reality of analytics in the world

Adam J. Ducey, IBM, Hazelwood, Missouri; Nigel Guenole, IBM Smarter Workforce Institute, London, United Kingdom; Sara P. Weiner, IBM Smarter Workforce, Tucson, Arizona; Hailey A. Herleman, IBM Smarter Workforce, Frisco, Texas; Robert E. Gibby, IBM, Cincinnati, Ohio; Tanya Delany, IBM, Milan, Italy.

The opinions expressed in this article are those of the authors and do not necessarily reflect the views of IBM's positions, strategies, or opinions.

Correspondence concerning this article should be addressed to Adam J. Ducey, IBM, 325 James S. McDonnell Boulevard, Hazelwood, MO 63042. E-mail: aducey@us.ibm.com

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

of big data. We enthusiastically accept their invitation to share our perspective that extends the discussion in three key areas of the focal article—that is, big data sources, logistic and analytic challenges, and data privacy and informed consent on a global scale. In the subsequent sections, we share our thoughts on these critical elements for advancing I-O psychology's role in leveraging and adding value from big data.

Big Data Is About Moving Beyond Traditional Data Sources

Although we agree with the authors that big data is characterized by volume, variety, and velocity, we believe that the definitions offered could be extended to paint a more complete picture of how the world of big data is different from our traditional world of data analysis. There are new frontiers in the variety of big data, beyond linkage analysis and discrete data sets, available to I-O psychologists. As noted by Zikopoulos, Eaton, deRoos, Deutsch, and Lapis (2012), 80% of the world's data are unstructured or semistructured. This includes sources such as videos, pictures, audio files, free text fields, presentations, word processing documents, e-mail messages, sensors, radio-frequency identification chips, and click streams. As a result, much of the current thinking about big data in I-O psychology focuses on the 20% of data that are easily accessed in relational databases. In order to truly harness the power of big data, however, we must expand our conceptualization of the variety of data and look at these novel sources.

Work is already underway to develop methods to store, access, and analyze this unstructured information (e.g., Ferrucci & Lally, 2004). In addition, the rate of data accumulation is increasing naturally in businesses today and for workforce data specifically. Organizations are struggling to gain insight from the data as fast as they are created. Recent estimates project that 44 zettabytes (44 trillion gigabytes) of data will be created by 2020, an increase of 10 times from 2013 (EMC & IDC, 2014). Along with the velocity of data in other aspects of our lives outside of work, this has created an expectation for very quick or even real-time insights from the data we create. When we go to pay for a list of items purchased from a website, a customized list of recommended additional purchases is instantly waiting for us. The data are there and accumulate quickly. Our expectation that organizations will also create value from data about employees quickly or even instantly is very real.

Another characteristic of big data has been offered by numerous authors (e.g., Ryan & Herleman, in press). Veracity describes big data as characterized by uncertainty. There are tons of missing data, many areas where data are created that have no real value or meaning, and other aspects that make the data very difficult to understand or interpret. Our more traditional data analysis techniques emphasize cleaning datasets, imputing missing values,

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

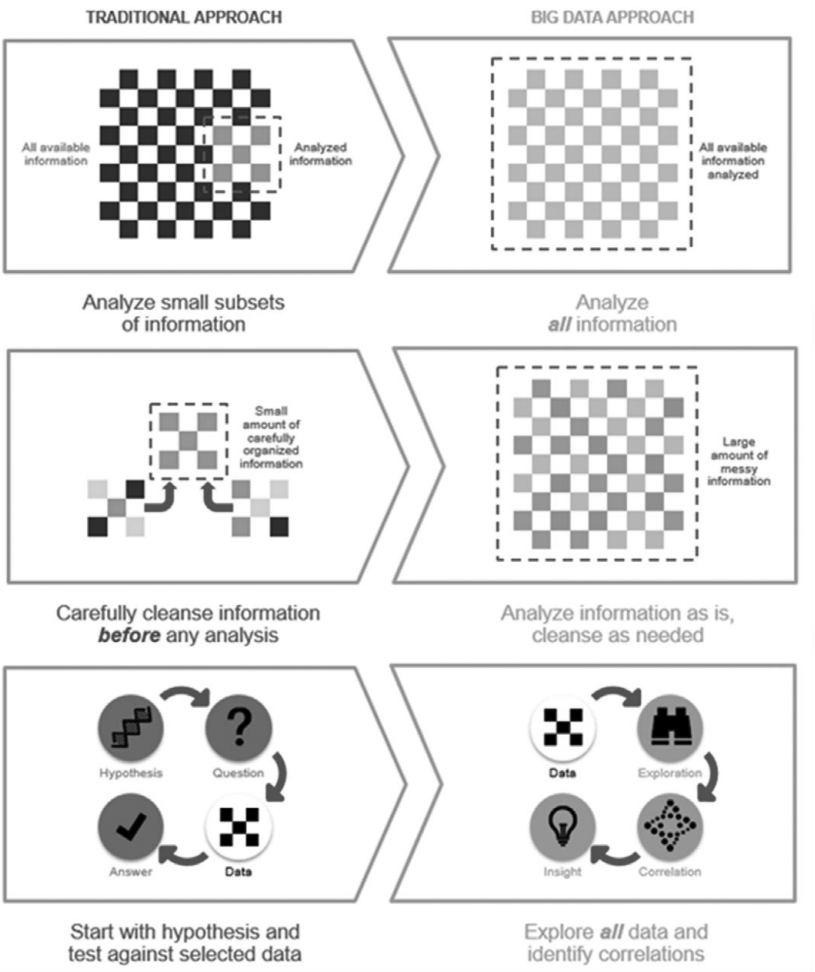

Figure 1. Contrasting traditional and big data approaches (from Ryan & Herleman, in press).

and collecting complete data in controlled settings whenever possible. Big data challenges this way of thinking and asks us to consider how we change our methods to gain valuable insights from very imperfect data. Big data is not just about a large data set, it is asking us to filter petabytes of data per second from almost any connected device, analyzing the data while still in motion, deciding what if any data must be stored, and even using analytics tools to virtually integrate the data with data stored in traditional warehouses. See Figure 1 for an illustration of these differences (adapted from Ryan & Herleman, in press).

It is clear that opportunities exist for I-O psychologists to partner with computer and data scientists, statisticians, and others to learn from and apply methods to incorporate the ever increasing and imperfect unstructured

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

talent data. These new forms of data arise from nontraditional data sources such as social learning and collaboration tools on corporate intranets and social media platforms that are becoming more common in organizational research. The distinct value of I-O psychology in this new world of big data is in providing a behavioral science and theoretical overlay for the data considered, analyses used, insights drawn, and creation of ongoing processes and systems leveraging big data to inform decisions on talent in the workplace. The challenge is that we have to understand enough about the new methods of data collection, management, and analysis to be able to partner effectively in the effort.

Big Data Creates New Logistic and Analytic Challenges

Expanding the realm of big data creates unique logistic and analytic challenges not discussed in the focal article. Our intent is not to provide an exhaustive presentation of these issues but rather to highlight those most salient to practitioners. For a more comprehensive discussion of these and other challenges related to big data, see Ryan and Herleman (in press).

As has already been established, big data is potentially high value, varied in form, accumulating quickly, and highly flawed in many cases when compared with our historical expectations. As a result, the traditional research paradigm of extracting the data and analyzing them over multiple days or weeks and then communicating the findings out to a broader audience is, in some cases, too slow. Today, organizations expect big data to monitor data sources in real time to identify patterns as they occur, provide regular updates, and generate new insights. Given this environment, there is a need for I-O psychologists to rapidly upskill their capabilities in data platforms, logistics, and analytics to better handle these demands.

In addition, there are myriad questions to consider. Where are your employee data now, and how can you retrieve them? Once you have them all, where will you put them? How will you update them? How will you keep them secure? Once you merge and store them, how do you retrieve them for analysis? How long do you keep them before moving them somewhere else for archiving? Many organizations find that their employee data are housed in many different servers and systems, both internal and external to the organization. Upon investigation, many organizations find that in order to get a complete picture of everything they know about employees, there are many protocols and sometimes costs associated with retrieval.

Also, a snapshot of data may not be sufficient for answering a research question. Organizations need to set up data feeds or connections that allow for real-time updates and insights from these many data sources. Organizations are wrestling with where to store the data and what tools are required

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

to access them for analysis. Are the same methods used in marketing and finance applicable to employee data? As addressed in the focal article, organizations are still sorting through how the protocols should vary depending on data type, subject matter, source, and country in order to protect individual rights and organizational security.

On Big Data Methods

In addition to the platform and logistic concerns, there is the matter of analyzing huge amounts of structured and unstructured data to generate insights and value for business. The challenge is that machine learning, cognitive computing, linguistic analysis, and other methods employed by statisticians, data scientists, and related disciplines are complex and typically not part of the curriculum in I-O psychology programs. These analysis techniques go beyond our traditional regression and modeling approaches for managing structured and discrete sets of data. There exists, however, a subset of quantitatively focused I-O psychologists who do have capability in these analytic methods. I-O big data experts (e.g., Oswald & Putka, 2015) can serve to bridge the gap across I-O theory, big data analytics, and business needs. A primary example offered by the focal article is linkage analysis where machine-learning techniques have been employed by I-O psychologists (e.g., Gibby, McCloy, & Putka, 2013).

These big data logistic and analytic requirements can get complicated quickly, and we are not suggesting that I-O psychologists need to become leading experts in data infrastructure, data platforms, software engineering, international law, and cybersecurity. However, we do suggest that I-O psychologists need to understand these logistical and analytic methods, ask the right questions, and bring in the right experts at the right time, including recognizing when the tried and true methods may still be the most appropriate. We also recommend comparing multiple methods, traditional statistical methods and machine-learning methods, to identify the most relevant approaches. In addition, consider that new data analysis methodologies (e.g., decision trees, random forests) must be applied to big data or that using big data is even the best approach.

Big Data at Any Cost?

We have sometimes found it better to spend resources collecting a "small" new data set of high quality (e.g., randomly sampled) that answers your question precisely rather than integrating a big data set from disparate sources that is of lower quality and that might not precisely answer your question. This is a common issue we see businesses grappling with—that is, proceeding with resource intensive integration of big data sets from disparate sources to find there are very few "overlapping data fields," the data are low

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

quality, or both. In order to decide whether integration of existing data sets is preferable to collection of new data, I-O psychologists need to be very clear about what data they require to provide a satisfactory answer to the question at hand. We recommend answering this question as formally as possible to help organizations decide their preferred course. Inevitably, the aforementioned logistic complexities will arise if the decision is to integrate existing data sources. Many of these issues are better addressed from the outset when data are collected. If they are not or cannot be satisfactorily resolved, it is sometimes better to collect new "small" data that will more precisely answer your question, be less costly, and be faster.

Data Privacy and Informed Consent Still Matter in the Era of (Truly) Big Data

I-O psychologists need to develop an understanding of how to maximize the discoveries from big data while also protecting the individuals' rights and organizations' security. Guzzo et al. provide a discussion of many useful data privacy protection strategies. These approaches are important for helping to protect individuals and organizations from data breaches while simultaneously allowing analysts to generate insights from these massive and varied datasets. However, there is a trade-off to manage between privacy and discovery. For example, sharding, or partitioning the data into smaller datasets, can be counterproductive to the very goals of big data analytics. Sharding is associated with a traditional approach to analytics in that you generate a priori hypotheses, clean your data, and analyze small subsets of information based on your specific research questions. This approach is in contrast to a truly big data approach that analyzes all information, cleans data as needed, and explores all data simultaneously to identify meaningful relationships (Ryan & Herleman, in press). Although the intent of sharding is to protect individuals, in practice it may limit data insights. As a result, we believe that organizations should consider alternative privacy strategies. That said, we have always operated within clear rules with respect to privacy protection for individuals and determining meaningful insights; these considerations must still prevail and will serve us well in this new realm.

In addition, we believe there is benefit in extending the data privacy requirements presented in the focal article. The data privacy plan focuses only on the analyst and the data. When working with big data, both direct and indirect access to the data should be part of the data privacy plan. Indirect access refers to access by anyone acting in a support capacity. When dealing with big data of a personal or sensitive manner, it is now our responsibility to anticipate indirect access and ensure protections. These protections also need to cover partners/vendors who will have direct or indirect access to the data.

Regarding informed consent, we echo Guzzo et al. in that recommendations provided by the Society for Industrial and Organizational Psychology

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

and the American Psychological Association leave gray areas in the context of big data, especially as the guidelines were designed for research mostly in universities rather than organizational settings and for a time when studies were small and localized. However, the Internet and digital age have spawned a whole new generation of researchers working with big data and in international collaborations. This means that researchers can increase the output of their work in collaboration rather than in competition, ensuring that research data are used effectively and efficiently.

A key question that should be discussed further is what types of unstructured and readily available public data sources should be exempt from informed consent? In what cases would there be implied informed consent? Current guidelines provide clear language that you must obtain informed consent when you use data sources such as individuals' photos and videos in research settings, but it is unclear whether informed consent captured by one organization (e.g., Facebook) is sufficient when a secondary organization uses publicly available data for other purposes. Data scientists are already turning to "found" data, creating new challenges for thinking about how to achieve consent or how to think ethically about the people behind those data. Working with big data provides us the opportunity and challenge to merge data that were collected with consent for one purpose and easily repurpose them for other business/research questions. How do we get permission without knowing all future uses of the data?

In addition to adhering to the Society for Industrial and Organizational Psychology and American Psychological Association guidelines, we recommend incorporating data privacy and security protocols reflecting the regulatory and legal requirements at a country level. There are more than 50 international legal and industry mandates focused on data privacy. These requirements, along with industry or practice-based guidelines, must be taken into consideration when working with big data.

Consider the distinctions among the psychological meaning of *privacy*, *security*[1] in a technological sense, and *compliance* in a legal sense. *Privacy* in the psychological sense is widely agreed to refer to an individual's ability to regulate how much information about the self is known to others (e.g., Westin, 1967). *Security* in a technological sense refers to methods for preventing disclosure of sensitive information to unintended recipients, a field sometimes referred to as privacy enhancing technologies (PET: Navarro-Arribas & Torra, 2015). *Compliance* in a legal sense refers to whether an organization's policies and procedures conform to the requirements of relevant legislation (e.g., Herrmann, 2007). I-O psychologists working with big data

[1] Security is also referenced in big data discussions in the context of issues of homeland defense—for example, how much privacy individuals are willing to sacrifice to ensure safety from terrorism. The issues most I-O psychologists deal with are intraorganizational, and as a result, we do not discuss this interpretation of the term *security* here.

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

must address all three areas. However, meeting the demands of one of these areas may not meet the requirements in others.

Though the focal article authors provided recommendations with "no implication of required compliance," we believe I-Os are truly the stewards of the data we collect and/or access. We have an obligation to protect the privacy of those data. This includes educating and leading in the field of big data analytics to guide data scientists, software engineers, and others who are now involved in work traditionally in our domain.

Conclusion

The big data revolution and the emerging field of people analytics give us cause to be excited as I-O psychologists. We believe the expertise and insights shared by Guzzo and colleagues provide a first step toward establishing a way forward for I-O psychologists to effectively operate in this new era. We also need to quickly advance our thinking about what big data means for our field. The aim in sharing our thoughts about expanding the definition of big data and the associated logistic, analytic, and privacy challenges is to ensure that our profession can lead the way and effectively partner with other disciplines to add value as the evolution of gleaning insights from big data continues.

References

EMC, & IDC. (2014, April). *The digital universe of opportunities: Rich data and in the increasing value of the Internet of things*. Retrieved from http://www.emc.com/leadership/digital-universe/2014iview/index.htm

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Journal of Natural Language Engineering, 10*(4), 327–348.

Gibby, R. E., McCloy, R. A., & Putka, D. J. (2013, April). *Viewing linkage research through the lenses of current practice and cutting-edge advances*. Workshop conducted at the 30th Annual Conference of the Society for Industrial Organizational Psychology, Houston, TX.

Guzzo, R. A., Fink, A. A., King, E., Tonidandel, S., & Landis, R. S. (2015). Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*(4), 491–508.

Herrmann, D. S. (2007). *Complete guide to security and privacy metrics: Measuring regulatory compliance, operational resilience, and ROI*. Boca Raton, FL: Auerbach.

Navarro-Arribas, G., & Torra, V. (2015). Advanced research on data privacy in the ARES Project. In G. Navarrio-Arribas & V. Torra (Eds.), *Advanced research in data privacy* (pp. 3–14). Cham, Switzerland: Springer International.

Oswald, F. L., & Putka, D. J. (2015). Statistical methods for big data. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge. Retrieved from http://www.researchgate.net/publication/271836790_Statistical_methods_for_big_data_A_scenic_tour

Ryan, J., & Herleman, H. A. (in press). A big data platform for workforce analytics. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge.

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press

Westin, A. F. (1967). *Privacy and freedom*. New York, NY: Atheneum.
Zikopoulos, P. C., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2012). *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. New York, NY: McGraw-Hill.

Conducting Ethical Research With Big and Small Data: Key Questions for Practitioners

Kathryn Dekas
Google, Mountain View, California

Elizabeth A. McCune
Microsoft, Redmond, Washington

The focal article (Guzzo, Fink, King, Tonidandel, & Landis, 2015) sought to "raise awareness and provide direction with regard to issues and complications uniquely associated with the advent of big data" (p. 492), and we commend their success in offering Society for Industrial and Organizational Psychology (SIOP) members a solid foundation and resources on which to draw. Our aim here is to extend their position, particularly to drive the conversation toward concrete recommendations for how industrial and organizational psychologists (I-Os) working in industry can apply the principles set forth in the focal article in our day-to-day work, specifically around the issue of avoiding ethical missteps in this new landscape.

Our ideas described below are the product of a working group assembled prior to the SIOP 2015 conference in preparation for a panel discussion titled "Guidelines for Ethical Research in the Age of Big Data" (McCune et al., 2015). The working group included four I-Os working in the tech, retail, and consumer product goods industries; an employee data privacy expert from the tech industry; an associate director of an institutional review board (IRB) at a top university; and a member of a European Works Council.

The original aim for the panel was to provide SIOP session attendees with the proverbial "dos" and "don'ts" list in conducting ethical research with big data to help newcomers to the big data/data science world engage with these new methods in an ethically sound way. However, as we worked through the process it became increasingly clear that issues of ethics around the use of data—big or small—are highly subjective and context dependent,

Kathryn Dekas, Google, Mountain View, California; Elizabeth A. McCune, Microsoft, Redmond, Washington.

Correspondence concerning this article should be addressed to Kathryn Dekas, Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043. E-mail: kdekas@google.com

https://doi.org/10.1017/iop.2015.83 Published online by Cambridge University Press