The Spanish Journal of Psychology (2020), **23**, e44, 1–5.
© Universidad Complutense de Madrid and Colegio Oficial de Psicólogos de Madrid 2020
doi:10.1017/SJP.2020.45

CAMBRIDGE UNIVERSITY PRESS

How Can Big Data Science Transform the Psychological Sciences?

Betsy H. Albritton and Scott Tonidandel

University of North Carolina at Charlotte (USA)

Abstract. Big data and related technologies are radically altering our society. In a similar way, these approaches can transform the psychological sciences. The goal of this commentary is to motivate psychologists to embrace big data science for the betterment of the field. Big data sources, algorithmic methods, and a culture that embraces prediction has the potential to advance our science, improve the robustness and replicability of our research, and allow us to focus more centrally on actual behaviors. We highlight these key transformations, acknowledge criticisms of big data approaches, and emphasize specific ways psychologists can contribute to the big data science revolution.

Received 30 July 2020; Revised 10 September 2020; Accepted 29 September 2020

Keywords: algorithms, big data, data science, deep learning, machine learning

Big data is defined in numerous ways, but it is often discussed in the context of several Vs; the most common of which are: *volume* or quantity of data, *variety* or diversity of data, and *velocity* or the speed of data acquisition (Laney, 2001). Big data methods strive to generate models from these high dimensional datasets that achieve enhanced predictive accuracy while also managing large sets of predictors and avoiding overfitting the model to the data. In recent decades, the conversation surrounding big data has grown (e.g., Tonidandel, et al., 2015; Woo, et al., 2020), but the number of empirical studies using big data methods in the psychological sciences remains low. Despite journals dedicating special issues to discussing big data, the social sciences are not as research active in embracing and applying these modern methods (Harlow & Oswald, 2016).

The purpose of this commentary is to combat this apparent resistance in the hopes of energizing psychologists to explore both old and novel research inquiries with the use of big data methods. To achieve this end, we will describe what it would look like if psychologists were to embrace a big data culture of algorithmic predictions. We will emphasize what we see as some of the key transformations to psychological science that are possible by making this transition and challenge prevailing assumptions that limit the acceptance of big data methods in psychology. Finally, we will highlight the vital role psychologists can play in the big data science revolution due to their strong background in measurement, experience with the ethical collection and use of human subjects data, and broader training in the psychological sciences.

Embracing Prediction

Breiman (2001), in his seminal article, describes two approaches to analyzing data. The first of these approaches, called the data modeling culture, emphasizes a deductive approach where a model is hypothesized and then validated based on goodness-of-fit. This is the dominant paradigm in our discipline where psychologists explore causal relationships between constructs to explain human behavior using controlled experimental research designs and statistical techniques such as ANOVA, regression, and structural equation modeling. Though psychologists' infatuation with the data modeling culture has certainly been useful, their sole reliance on it has likely stifled scientific progress. Many of the models psychologists test using this paradigm are likely oversimplified and underdeveloped, and often, competing models, that may fit the data just as well, are never tested nor evaluated. Thus, they may be placing their faith in models that fit the data, but nevertheless are not true.

Breiman's (2001) alternative to the data modeling culture is the algorithmic modeling culture, which encapsulates the approach of big data methods and

Correspondence concerning this article should be addressed to Betsy H. Albritton. University of North Carolina at Charlotte. 28223–0001 Charlotte, North Carolina (USA).

E-mail: ealbritt@uncc.edu

Conflicts of Interest: None

Funding Statement: This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

How to cite this article:
Albritton, B. H., & Tonidandel, S. (2020). How can big data science transform the psychological sciences? *The Spanish Journal of Psychology, 23.* e44. Doi:10.1017/SJP.2020.45

https://doi.org/10.1017/SJP.2020.45 Published online by Cambridge University Press

focuses on maximizing predictive accuracy. Instead of applying what we think we know about the data, algorithmic models get to learn the qualities of the data. Such an approach confers a number of critical advantages like the ability to incorporate a wider variety of variables and more complex relationships between variables, even without knowing a priori what exactly the form of those relationship should be. By emphasizing predictive accuracy and capturing more elements of the relationship between variables, one ultimately learns more about the problem under investigation. For research to progress, psychology needs more flexible, complex models and the algorithmic culture enables this to happen. Thus, we argue that psychology as a science needs to embrace the algorithmic culture of big data methods and by doing so, we have the opportunity to transform the psychological sciences.

Transforming the Psychological Sciences

Returning to Our Roots and the Study of Behaviors

Psychology is the science of behavior, but a critical appraisal of the field unfortunately reveals that psychologists do little to study actual behavior. In 2007, Baumeister and colleagues lamented that psychology has become the study of self-report after noting that, in an entire issue of the *Journal of Personality and Social Psychology*, none of the 38 studies contained a direct observation of behavior. Fast forward a decade and a half and the situation remains just as bleak. Banks, et al. (2020) coded 184 primary samples appearing in five top journals in applied psychology and found that only 3% of the variables studied in these articles were behavioral in nature. The apparent neglect of actual behaviors may be attributable to a variety of causes including the ease of which self-report data can be obtained (Day, 2014) and the perceived difficulty or seeming infeasibility of collecting behavioral data (Baumeister et al., 2007). Fortunately, big data methods may offer a solution to these barriers and enable the field to return to a more central focus on actual behaviors. Wearable sensors now permit the collection of highly granular continuous streams of behavior cataloguing all sorts of relational behaviors (Matusik, et al., 2019). Advances in video analytics permit the analysis of real time human behavior via personal wearable cameras (Blake, et al., 2020) or via the ever growing network of cameras that exist in our homes, at our schools, at our places of work, and in public spaces in cities throughout the world (Aghajanzadeh, et al., 2020). Technological advances in the capture and analysis of real time behavior data have broken down the preexisting barriers that have prevented psychological research from focusing more on behaviors.

Addressing the Replication Crisis

The entire field of psychology currently faces a replication crisis as our field routinely struggles to alleviate worries surrounding an inability to replicate previously published findings. In a study conducted by Hartshorne and Schachner (2012), only 49% of psychological results were fully replicated out of 257 attempts. Ioannidis (2005) demonstrated the prevalence of the replication crisis across scientific fields, and yet there is still no explanation for why psychology elicits more fraudulent practices than other fields of science (Carpenter, 2012). Some of the commonly identified culprits are behaviors such as p-hacking (Simmons, et al., 2011), HARKing (Kerr, 1998), questionable research practices (QRPs; Banks, et al., 2016), and a lack of emphasis on cross-validation (de Rooij & Weeda, 2020).

How can big data methods help to improve the robustness of psychological research? Grand et al. (2018) describe several elements of a robust science that, upon closer inspection, are natural features of big data methods. Most psychological studies suffer from small underpowered samples, and publication pressures incentivize QRP's like seeking out statistically significant results and crafting hypotheses post hoc. Big data methods, in contrast, tend to have large samples making statistical significance testing superfluous, thereby eliminating the need for p-hacking. Big data methodology also emphasizes cross-validating models, which would alleviate some of the symptoms of the replicability crisis. Traditional methods are predominantly concerned with how well the model fits the present sample, but predictive models and big data methods evaluate the quality of a model based on how accurate it is in making predictions on a new sample of data (Shmueli, 2010). This rigorous process of model building helps guard against overfitting by minimizing the likelihood of capitalizing on noise of randomness in a particular study's dataset and increase the likelihood of finding similar results in a new sample.

Advancing our Science via Prediction Models

Our current theoretical models are limited in scope. Theoretical models in psychology today are often static, struggle with a large number of predictor variables, and cannot handle complex nonlinear associations between variables (Tonidandel, et al., 2018). For years, psychologists have expressed the desire to make time a more central feature in their models. Methodological advances such as event studies, growth curve modeling, and regression discontinuity designs have helped to propel the field forward. However, studies using these methods still tend to be singular snap-shots of a phenomenon over time and rarely is one able to collect and analyze a continuous stream of data repeatedly.

https://doi.org/10.1017/SJP.2020.45 Published online by Cambridge University Press

However, velocity is a key feature in big data research that permits the continuous evaluation of a model's predictive capabilities allowing for models to change as new data comes in. When the COVID–19 pandemic struck, fields such as finance that rely heavily on predictive modeling were able to quickly recognize that their existing models were no longer applicable and adjusted to the new reality (Knight, 2020). If psychologists had leveraged similar algorithmic modeling techniques prior to COVID–19, the field would be able to react much more quickly recognizing which models should now be considered outdated and would be able to make changes accordingly.

Traditional methods in psychology can typically only accommodate a limited number of variables. As a result, psychological models are often overly simplistic, including very few variables and examining mostly linear relationships between variables. Unfortunately, this has constrained the field's ability to accurately predict and explain the complexity that is human behavior. In contrast, big data methods can uniquely capture complexity in psychological constructs and phenomenon that are otherwise unexplored by current techniques. Big data methods are ideally suited for handling high dimensions data (i.e., lots of variables), and various modeling approaches can identify both linear and nonlinear relationships without having to specify those relationship a priori. As a result, these methods can help to identify key variables and/or complex relationships that may currently be missing from existing models.

Tempering Algorithmic Concerns

An acceptance of the algorithmic modeling culture in psychology is not without limitations. To start, psychologists lack training in and exposure to many predictive methods. Additionally, the psychological sciences stress a theory-driven approach to modeling and view algorithmic approaches as atheoretical, uninterpretable, and biased. Though each of these concerns contains a kernel of truth, these concerns are often exaggerated and may in fact be more myth than reality.

Anderson (2008) declared that the end of theory was at hand because big data makes the scientific method obsolete. Unfortunately, hyperbolic proclamations such as these create the mistaken impression that big data science is atheoretical, which is patently false. It is true that big data science is less theory driven than our traditional hypothetico-deductive approach, but big data approaches are certainly theory-laden and perhaps can be best described as theory-informed (Haig, 2020). The acceptance and implementation of algorithms in the psychological sciences does not inherently mean that psychologists would disregard the use or importance of theory. In fact, theory plays a critical role in the creation and interpretation of predictive models. A main consideration in big data and algorithmic modeling is the choice of predictors and which features to engineer. Theory is critical in guiding the feature engineering process, and the inclusion of theoretically relevant variables in the model is a necessity to have highly interpretable and meaningful results (Urban & Gates, 2019).

Algorithmic approaches are commonly viewed as black boxes. The vast number of predictors utilized in big data projects and the complexity of algorithmic models (e.g., artificial neural networks) make big data approaches substantially harder to interpret than simpler analytical approaches (e.g., ordinary-least-squares regression). Although there is some truth to this criticism, it neglects several important facts. The first is that accuracy should be of paramount importance for any scientific model. A simple, easily explainable model that produces inaccurate predictions is not useful. One would be better served having an accurate model and trying to interpret it in all of its complexities. A model of our universe where Earth is at the center is certainly simpler than one that has our galaxy as but one of several hundred billion; nevertheless, we are better served by the more complex and harder to explain model. A second shortcoming of this black box criticism is that it neglects an ever growing field of research on explainable artificial intelligence (AI). When big data methods were in their infancy, lamenting about the interpretability of these models was perhaps well-founded. Now, however, tools exist that allow researcher to understand what features of a complex algorithm are driving the predictions being made (see Local Interpretable Model-Agnostic Explanations [LIME], Ribeiro, et al., 2016 as an example). In fact, one can even point to instances where algorithms can be even more revealing than traditional methods. For example, when humans make decisions or provide ratings, they may lack insight into what features are responsible for the response they are providing. In contrast, one can use one of these explainable AI tools to investigate why the algorithm is behaving the way that it is. In some ways, an algorithm that is trained on human labeled data can allow researchers to peer into the black box of the human mind to begin to infer why the human rater behaved the way they did (Thompson, et al., 2020).

There is also a pervasive belief that algorithmic approaches are inherently biased, and numerous high profile examples exist to support this perception (e.g., Dastin, 2018; O'Neil, 2016). Despite these spectacular failures, this criticism is not really one that should be leveled at the algorithms but rather at the underlying data. Existing societal injustices are baked into the data that the algorithms train on. The algorithms just learn

https://doi.org/10.1017/SJP.2020.45 Published online by Cambridge University Press

and replicate the bias that already exists. Thus, these approaches are no more biased than the current non-algorithmic approaches. More importantly however, algorithmic researchers can inspect the algorithms for bias and can make efforts to train that bias out of predictions. It may even be easier to train bias out of algorithmic predictions than to train human decision makers to ignore their own subconscious biases.

Psychologists' Key Role in the Revolution

One does not need to be a data scientist or a computer scientist to play an integral role in the big data science revolution. Even with the above transformations to psychological research, psychologists can advance big data methodology itself due to their extensive training and knowledge of measurement. Additionally, psychologists frequently face ethical considerations regarding the collection and security of human-generated data, which makes them uniquely suited to handle datasets that capture human behaviors and processes.

More important than the utilization of big data is the appropriate measurement and collection of big data, and psychologists' extensive training in measurement can greatly improve the current practices in big data research (Woo, et al., 2020). Though data scientists are extremely capable at applying the latest and greatest algorithmic tools to the data that is made available to them, their lack of training in psychometrics can sometimes lead to situations where big data projects are criticized as "garbage in, garbage out". Measurement is essential to maximizing the returns from big data projects, and psychologists, with decades of scientific literature and years of focused graduate training on how to accurately measure constructs of interest, are ideally positioned to contribute. Although data scientists have access to large quantities of data and the amount of data generated by humans every day continues to grow, the amassment of large datasets does not automatically result in more accurate or nuanced results. Jacobucci and Grimm's (2020) study demonstrates how impactful measurement is to the big data literature. Jacobucci and Grimm (2020) tested the influence of measurement error on model estimation in a simulation study and demonstrated that algorithmic models perform really well when the data is exceptionally clean and measured without error. Nevertheless, when error was introduced to the data, $?^2$ was reduced by 76% in big data models. With the introduction of the slightest bit of measurement error, predictive accuracy falls. Thompson, et al. (2020) provide additional evidence of the meaningful gains that can be made when measurement is a central component of a big data study.

Psychologists also have a key roles to play in ensuring the ethical use of big data. Psychologists have advanced training and familiarity with rules governing the protection of human subjects stemming from the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). The central principles of *Respect for Persons*, *Beneficence*, and *Justice*, continue to be essential for ensuring ethical use of big data sources (Paxton, 2020). These principles are imperative for the current use of big data, but also for unforeseen circumstances. As the volume of big data sources rises, new challenges concerning the ethical and fair use of that data will also increase and require the creation and enforcement of new privacy and security measures (Guzzo, et al., 2015). Psychologists, by virtue of their history and training, can help to ensure that the rights and dignity of participants are protected.

Although we strongly espouse that the field of psychology embrace these big data methods, we do not want to be accused of big data hubris (Lazer, et al., 2014), and we recognize that limitations exist for these methods as they do with any method. Nevertheless, we believe that big data, associated algorithmic methods, and a culture that embraces prediction affords an opportunity to transform the psychological sciences and propel the field forward.

References

Aghajanzadeh, S., Jebb, A. T., Li, Y., Lu, Y.-H., & Thiruvathukal, G. K. (2020). Observing human behavior through worldwide network cameras. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 109–123). American Psychological Association. https://doi.org/10.1037/0000193-006

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Banks, G. C., Woznjy, H. M., & Mansfield, C. (2020). *Where is "behavior" in theories of organizational behavior? A call for a behavioral revolution* [Manuscript submitted for publication]. Belk College of Business, University of North Carolina at Charlotte.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323–338. https://doi.org/10.1007/s10869-016-9456-7

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self- reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Blake, A. B., Lee, D. I., De La Rosa, R., & Sherman, R. A. (2020). Wearable cameras, machine vision, and big data analytics: Insights into people and the places they go. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 125–143). American Psychological Association. https://doi.org/10.1037/0000193-007

https://doi.org/10.1017/SJP.2020.45 Published online by Cambridge University Press

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558–1560. https://doi.org/10.1126/science.335.6076.1558

Dastin, J. (2018, October 9). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Day, D. V. (2014). The future of leadership: Challenges and prospects. In D. V. Day (Ed.), *Oxford Library of Psychology. The Oxford handbook of leadership and organizations* (p. 859–867). Oxford University Press.

de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Grand, J. A., Rogelberg, S. G., Allen, T. D., Landis, R. S., Reynolds, D. H., Scott, J. C., Tonidandel, S., & Truxillo, D. M. (2018). A systems-based approach to fostering robust science in industrial-organizational psychology. *Industrial and Organizational Psychology*, *11*, 4–42. https://doi.org/10.1017/iop.2017.55

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*, *8*, 491–508. https://doi.org/10.1017/iop.2015.40

Haig, B. D. (2020). Big data science: A philosophy of science perspective. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 15–33). American Psychological Association. https://doi.org/10.1037/0000193-002

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447–457. https://doi.org/10.1037/met0000120

Hartshorne, J., & Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience*, *6*, Article 8. https://doi.org/10.3389/fncom.2012.00008

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Knight, W. (2020, July 19). Even the best AI models are no match for the coronavirus. *Wired*. https://www.wired.com/story/best-ai-models-no-match-coronavirus/

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6* (70), 1.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343* (6176), 1203–1205. https://doi.org/10.1126/science.1248506

Matusik, J. G., Heidl, R., Hollenbeck, J. R., Yu, A., Lee, H. W., & Howe, M. (2019). Wearable bluetooth sensors for capturing relational variables and temporal variability in relationships: A construct validation study. *Journal of Applied Psychology*, *104*(3), 357–387. https://doi.org/10.1037/apl0000334

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. U.S. Department of Health and Human Services. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Paxton, A. (2020). The Belmont Report in the age of big data: Ethics at the intersection of psychological science and data science. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 347–372). American Psychological Association. https://doi.org/10.1037/0000193-016

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366. https://doi.org/10.1177/0956797611417632

Thompson, I., Koenig, N., Mracek, D., & Tonidandel, S. (2020). *Integrating deep learning and measurement science: Automating the subject matter expertise used to evaluate candidate work samples* [Manuscript submitted for publication]. Belk College of Business, University of North Carolina at Charlotte.

Tonidandel, S., King, E. B., & Cortina, J. M. (Eds.). (2015). *Big data at work: The data science revolution and organizational psychology*. Routledge. https://doi.org/10.4324/9781315780504

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525–547. https://doi.org/10.1177/1094428116677299

Urban, C., & Gates, K. (2019). *Deep learning: A primer for psychologists*. PsyArXiv. https://doi.org/10.31234/OSF.IO/4Q8NA

Woo, S. E., Tay, L., Jebb, A. T., Ford, M. T., & Kern, M. L. (2020). Big data for enhancing measurement quality. In S. E. Woo, L. Tay, & R. W. Proctor (Eds.), *Big data in psychological research* (pp. 59–85). American Psychological Association. https://doi.org/10.1037/0000193-004

Woo, S. E., Tay, L., & Proctor, R. W. (Eds.). (2020). *Big data in psychological research*. American Psychological Association. https://doi.org/10.1037/0000193-000

https://doi.org/10.1017/SJP.2020.45 Published online by Cambridge University Press