

CONTENTS

Contributors	vii
Introduction	3
<i>Sang Eun Woo, Louis Tay, and Robert W. Proctor</i>	
I. BACKGROUND AND OVERVIEW	13
1. Big Data Science: A Philosophy of Science Perspective	15
<i>Brian D. Haig</i>	
2. From Small-Scale Experiments to Big Data: Challenges and Opportunities for Experimental Psychologists	35
<i>Robert W. Proctor and Aiping Xiong</i>	
3. Big Data for Enhancing Measurement Quality	59
<i>Sang Eun Woo, Louis Tay, Andrew T. Jebb, Michael T. Ford, and Margaret L. Kern</i>	
II. INNOVATIONS IN LARGE-SCALE DATA COLLECTION AND ANALYSIS TECHNIQUES	87
4. Internet Search and Page View Behavior Scores: Validity and Usefulness as Indicators of Psychological States	89
<i>Michael T. Ford</i>	
5. Observing Human Behavior Through Worldwide Network Cameras	109
<i>Sara Aghajanzadeh, Andrew T. Jebb, Yifan Li, Yung-Hsiang Lu, and George K. Thiruvathukal</i>	
6. Wearable Cameras, Machine Vision, and Big Data Analytics: Insights Into People and the Places They Go	125
<i>Andrew B. Blake, Daniel I. Lee, Roberto De La Rosa, and Ryne A. Sherman</i>	

7. Human-Guided Visual Analytics for Big Data	145
<i>Morteza Karimzadeh, Jieqiong Zhao, Guizhen Wang, Luke S. Snyder, and David S. Ebert</i>	
8. Text Mining: A Field of Opportunities	179
<i>Padmini Srinivasan</i>	
III. APPLICATIONS	201
9. Big Data in the Science of Learning	203
<i>Sidney K. D'Mello</i>	
10. Big Data in Social Psychology	227
<i>Ivan Hernandez</i>	
11. Big Data in Health Care Delivery	255
<i>Mohammad Adibuzzaman and Paul M. Griffin</i>	
12. The Continued Importance of Theory: Lessons From Big Data Approaches to Language and Cognition	277
<i>Brendan T. Johns, Randall K. Jamieson, and Michael N. Jones</i>	
13. Big Data in Developmental Psychology	297
<i>Kevin J. Grimm, Gabriela Stegmann, Ross Jacobucci, and Sarfaraz Serang</i>	
14. Applying Principles of Big Data to the Workplace and Talent Analytics	319
<i>Q. Chelsea Song, Mengqiao Liu, Chen Tang, and Laura F. Long</i>	
IV. RECOMMENDATIONS FOR RESPONSIBLE AND RIGOROUS USE OF BIG DATA	345
15. The Belmont Report in the Age of Big Data: Ethics at the Intersection of Psychological Science and Data Science	347
<i>Alexandra Paxton</i>	
16. Promoting Robust and Reliable Big Data Research in Psychology	373
<i>Joshua A. Strauss and James A. Grand</i>	
17. Privacy and Cybersecurity Challenges, Opportunities, and Recommendations: Personnel Selection in an Era of Online Application Systems and Big Data	393
<i>Talya N. Bauer, Donald M. Truxillo, Mark P. Jones, and Grant Brady</i>	
18. Privacy Enhancing Techniques for Security	411
<i>Elisa Bertino</i>	
V. CONCLUDING REMARKS	425
19. Future Research Directions for Big Data in Psychology	427
<i>Frederick L. Oswald</i>	
Index	443
About the Editors	467

Introduction

Sang Eun Woo, Louis Tay, and Robert W. Proctor

We live in an exciting time for psychology. In this digital age, the power of technology enables us to collect and store a massive amount of data in a way that is faster, more plentiful, and more diverse than once imagined. These new data streams contain potentially useful information about human cognition, emotion, attitudes, and behavior that used to be prohibitively difficult—or even impossible—to capture using traditional research methods (Adjerid & Kelley, 2018). This limitation was due in part to costs but also a lack of technological infrastructure (e.g., smartphones, social media). Not only have technological advancements led to an abundance of new data streams, repositories, and computational power, they also have resulted in advances in statistical and computational techniques that have proliferated widespread analysis of such data in multiple domains (e.g., business, education, health care), improving our ability to predict psychological and societal outcomes. This is the era of *big data* for psychological research, which broadly refers to multiplying multiform data (e.g., structured, unstructured) and their supporting technological infrastructure (i.e., capture, storage, processing) and analytic techniques that can enhance psychological research (cf. Adjerid & Kelley, 2018; Harlow & Oswald, 2016).¹

¹Although we believe this is a broad enough definition that captures a reasonably large portion of “big data” discussions to date, it is important to note that there is little to no consensus on what the term *big data* exactly means; it is defined and conceptualized in a variety of different ways depending on the specific purposes and context of a given discussion.

<http://dx.doi.org/10.1037/0000193-001>

Big Data in Psychological Research, S. E. Woo, L. Tay, and R. W. Proctor (Editors)
Copyright © 2020 by the American Psychological Association. All rights reserved.

The rapid emergence of big data has been met with enthusiasm in many different fields—especially within applied settings. Although big data hold promise for psychology, there remains hesitancy and even skepticism about this approach. These attitudes stem not from Luddite mindsets but emerge from grounded concerns—apprehensions that arise from issues our field has had to tackle over its history. There are legitimate concerns about the loss of theoretical depth and specificity when big data advocates call for a moratorium on theory (e.g., Anderson, 2008). Psychologists are concerned with using data not only to maximize one's ability to predict meaningful outcomes but also to develop and further establish theories that explain the observed relationships. In this regard, more thoughtful and contextualized discussions are needed about what theoretical and scientific advancements can and cannot be achieved through the collection and analysis of big data. For example, an empirical finding that languages posted on social media can reliably predict the users' self-reported personality should not be immediately accepted as direct evidence for social media language's validity as a "measure" of personality; doing so would require philosophical and theoretical justification of the equivalence between prediction and causal explanation by personality constructs (for further discussions, see Woo, Tay, Hickman, & Saef, 2019).

Further, there are methodological and empirical questions as to the reliability, validity, and utility of big data for psychology. For one, the conditions under which the data are collected are largely uncontrolled (from an experimentalist's perspective, this is a major limitation). For another, there is limited evidence to date that machine learning algorithms that are currently used in applied and commercialized settings (e.g., recruitment and selection) can improve our ability to predict the outcomes of interest to psychologists (e.g., individual attributes that may be captured using traditional methods during selection, job performance, turnover; cf. Hickman, Tay, & Woo, 2019). Going beyond the prediction accuracy, another notable problem with using machine learning algorithms in hiring is the perpetuation of potential biases in the assessment and decision-making processes. For instance, Amazon had to halt their machine learning tool for sifting through applicant resumes when they observed that the algorithm had a preference for men over women (Dastin, 2018). These potential problems in control, prediction, and bias apply broadly to a vast array of psychological research and have to be better understood and clarified.

These questions occur at a time of deep self-reflection and self-criticism in psychology as psychologists work through the challenges of replication and reproducibility even in our well-established traditional research designs and methods. In this light, big data are intuitively appealing because the large swathes of data used have a greater statistical power to detect phenomena and yield more accurate statistical estimates. Yet, these same big data may introduce new issues or magnify old ones in replication and reproduction. These include aspects such as the way data are collected and cleaned, the type of hardware used, the expertise of the researcher, and the choice of analytic

techniques. Big data may also continue to underrepresent certain demographic groups, and analyses on subgroup differences may not be as robust.

Finally, there is the inevitable association of big data with “big brother” and digital surveillance. Where data are being collected at scale without informed consent for a specific research question, there are questions as to when and how these types of data can be used, to what extent informed consent is necessary, and how the privacy of individuals can be protected in the analytic and research process. The emergence of such data requires our field to work through ethical challenges of confidentiality and privacy as they pertain to big data.

Our desire to organize this handbook is twofold. We seek to showcase the opportunities of big data and its related methodologies for psychologists to study human behavior and cognition. At the same time, we believe that the key to unlocking this possibility requires addressing many of these concerns and challenges. For individual researchers to decide whether to incorporate some of these technology-driven ways of collecting and analyzing data into their existing methodological toolkits, there has to be a systematic and comprehensive understanding of big data in psychological research—both conceptually and methodologically. More broadly, our field requires a consolidated resource to continue informed conversations about the prospect of big data for psychology. This edited volume aims to address these needs.

In May 2018, we (the three editors of this book, Woo, Tay, and Proctor) organized a symposium, sponsored by Purdue University’s Department of Psychological Sciences, to which we invited 20 leading scholars and practitioners not only from different domains in psychology but also from other key related disciplines such as computer science, data science, philosophy of science, management, health care, and education. The goal of the symposium was to assess the current state of the big data movement critically in terms of its scientific value to the psychology community to address the conceptual, methodological, and ethical challenges and to identify areas in which future research is most needed. We sought to obtain a broad array of opinions on a wide variety of issues associated with the use of big data in psychology, which sparked a number of significant discussions and new insights. The current edited volume closely follows the structure and content of the symposium. Many of the symposium presenters have contributed a chapter to this edited volume, and we added several other authors to supplement the overall content coverage of the book.

As such, the overarching goal of this edited book is to provide readers with a bird’s-eye view of what big data mean for psychological research—new opportunities for collecting and analyzing psychological data, as well as challenges that may come with such possibilities. The book consists of four main parts, bookended by this introduction (by the editors) and the concluding chapter by Oswald (Chapter 19).

Part I, Background and Overview, features three chapters that discuss how big data have been (and should be) conceptualized and discussed in

psychology and other fields, addressing the “big picture,” conceptual questions from the perspectives of philosophy of science and research methodology. This section lays out the broad rationale for big data in psychology and discusses potential concerns of theoretical implications, research designs, and measurement reliability and validity.

In Chapter 1, Haig presents his scientific realist perspective (in terms of both global and local realism) on big data science, laying down a key philosophical foundation for understanding the role of science and scientific methods. He then explains inductive and abductive modes of science, in contrast with the hypothetico-deductive method, as alternative methodological paradigms for big data science. In particular, Haig proposes that the abductive method offers a useful framework for a broad array of big data inquiries because it guides not only the initial process of phenomena detection (on which the inductive method focuses) but also the subsequent efforts toward theory construction where researchers generate, develop, and appraise theories. Importantly, Haig warns against the antirealist claims that have been endorsed by some “big data enthusiasts” who completely dismiss the importance of theory and causal explanations. In doing so, he returns us to a more nuanced discussion of what constitutes a “theory” in the first place and points to the different interpretations of causation that exist in the literature.

In Chapter 2, Proctor and Xiong discuss how various aspects of big data present new opportunities and challenges for research in the experimental psychology tradition. They provide an impressive historical sweep of the experimental tradition and show how big data can contribute to experimental designs that have greater scale and representativeness. Critically, this enables research designs and analyses that would more directly account for environmental and contextual compatibilities and constraints on human cognition and behavior, providing a broader human–environment systems approach. This broader applicability has been one of the goals of experimentalists, and the big data approach can serve to complement existing approaches.

In Chapter 3, Woo, Tay, Jebb, Ford, and Kern offer an integrative overview of the different ways in which specific sources of big data (i.e., social media, wearable sensors, Internet behavior, public network cameras, smartphones) may be used for enhancing the measurement quality of psychological research in capturing affective and attitudinal states, personality traits, and interpersonal relationships. Specifically, the authors discuss issues of content relevance (as well as content deficiency and contamination), response processes, internal (factorial) structure, nomological net, and reliability as they relate to each of these data sources. In doing so, they provide psychologists with an analysis of the key considerations in reliability and validity for different big data sources.

Part II, *Innovations in Large-Scale Data Collection and Analysis Techniques*, contains five chapters that discuss tangible opportunities and strategies for using big data to enhance existing methods of data collection and analysis within psychology. These chapters also present specific challenges and future research directions.

In Chapter 4, Ford reviews the current state of psychological research that uses Internet behavior data. Specifically, he discusses major topics of interest covered in research to date on searches and page selections and views as captured by Google Trends and Wikipedia (i.e., physical health and disease; mental health; legal and illicit drugs; health behavior; consumer, economic, and financial market behavior; and policy and politics). He also presents some preliminary evidence for the criterion-related validity of these data in predicting psychologically meaningful outcomes. There is a substantial opportunity here for psychologists to collect and use such data for understanding and predicting behavior at the aggregate level.

In Chapter 5, Aghajanzadeh, Jebb, Li, Lu, and Thiruvathukal discuss the promise of using public network cameras for collecting observational data of human behaviors around the world. After reviewing the literature illustrating the value of using video data of public human behavior for psychological research, the authors introduce recent technological advances in computer science that enable researchers to access and analyze such video data. Specifically, CAM², a software platform built at Purdue University, can serve as a useful research tool for retrieving and analyzing data from worldwide network cameras, presenting psychologists with (largely unexplored) possibilities for capturing public human behavior on a large scale.

In Chapter 6, Blake, Lee, De La Rosa, and Sherman demonstrate the utility of wearable cameras in capturing the natural environments people encounter in their daily lives on a moment-to-moment basis. After providing a theoretical background for the importance of measuring situational cues and characteristics, the authors discuss the practical, legal, and ethical complexities of using wearable cameras in psychological research (e.g., privacy, obtrusiveness). In an illustrative example, the authors present a recent study that used wearable camera data analyzed with machine vision. This chapter clearly delineates the steps and protocols required for conducting such research.

In Chapter 7, Karimzadeh, Zhao, Wang, Snyder, and Ebert provide an overview of human-guided visual analytics that will be useful for psychological research using big data approaches. With the abundance and complexity of data, there is a need to visually present information in an interactive manner aligned with known perceptual and cognitive principles to enable more effective use and parameterization of big data for research and decision making. The implication is that advances in psychology (e.g., perception, information processing, decision making) can, in turn, contribute to this field of big data visual analytics. Further, the adoption of visual analytics in various application domains creates opportunities for psychological research on the use of such technologies in real-world scenarios and how such use modifies or is affected by human behavior.

In Chapter 8, Srinivasan describes the overall landscape of various text-mining approaches, organized into three broad categories: information extraction, information inferencing, and literature-based discovery. Given that a substantial amount of the work in psychology focuses on the use of lexicon approaches (e.g., dictionary approaches) in information extraction,

Srinivasan discusses the limitations of lexicons and how information extraction can be enhanced through template approaches that provide greater precision in identifying relevant sources of information. Several examples—including belief surveillance, personality perceptions, and life satisfaction—are used to illustrate this. This substantially more accurate approach requires psychologists to provide a priori content specification and operational definitions, highlighting the importance of construct and measurement validity.

Part III, *Applications*, contains six chapters that summarize how big data have been (and may be) used in specific subdisciplines of psychological sciences and other related fields. These areas include learning and education (Chapter 9), social psychology (Chapter 10), health care (Chapter 11), language and cognition (Chapter 12), developmental psychology (Chapter 13), and industrial and organizational psychology (Chapter 14).

In Chapter 9, D'Mello provides a compelling case for the promise of data-driven, technology-enabled approaches to advancing the science of learning. His examples of such approaches range from innovations in formative assessment techniques (e.g., Bayesian knowledge tracing, additive factor models) to various analytic techniques for dealing with transactional, sequential, linguistic, and multisensory-multimodal behavioral data in education research. Clearly, the field of educational research is embracing the promise and possibilities of big data.

In Chapter 10, Hernandez begins by briefly discussing how four defining characteristics of big data (i.e., volume, velocity, variety, veracity) may complement social psychological research. After extensively cataloging the sources of big data commonly used in social psychology (e.g., website messages, profile information, wearable sensors), he reviews the social psychology literature for key trends in big data research—organized by 12 focal topic areas of social psychology (e.g., attitudes, culture, diversity, motivation, groups). He also offers a summary of the linkages between key social psychological constructs and big data measures. This will serve as an important resource for not only social psychologists but also psychologists from different fields.

In Chapter 11, Adibuzzaman and Griffin note that the progress of big data research in health care has been slow, despite its significant promise. The authors offer helpful insights into reasons behind such delays, such as multi-layered issues of patient privacy in storing, managing, and sharing electronic health records and the limitations of drawing causality (as normally done in health care research through randomized controlled trials) from retrospective observational data. One implication here is that integrating and digitizing different threads of data for an entire field is a significant undertaking requiring careful consideration and coordination.

In Chapter 12, Johns, Jamieson, and Jones describe how big data approaches can be (and have been) applied to understanding human cognition, focusing on the case of building, testing, and validating computational models of natural language using large text corpora. They explain how the ability to store and analyze large amounts of data has enabled the development of better and more diverse sources of language characteristics. Johns and coauthors

highlight the fact that the precision provided by large-scale data sets allows researchers to test cognitive models at the level of individual items. The authors emphasize the importance of theory in guiding the scientific progress in this domain and suggest that theoretical insights gained from the abductive development of theories from big data must be accompanied by deductive verification under more controlled experimental conditions.

In Chapter 13, Grimm, Stegmann, Jacobucci, and Serang discuss the importance of longitudinal forms of big data in developmental psychology to understand growth. They focus on machine learning approaches for modeling explanations for growth trajectories in longitudinal panel data. These methods can help researchers explore and identify explanatory variables, which is especially useful in typical scenarios where there are too many variables for researchers to theoretically deduce the optimal set of explanations. The detailed explanation and application of these models to the Early Childhood Longitudinal Study—Kindergarten Cohort of 1998/1999 provide researchers a practical step-by-step guide to model selection and interpretation.

In Chapter 14, Song, Liu, Tang, and Long describe the rapid integration of big data into the field of industrial and organizational psychology to better understand and improve workplaces. There have been substantial advances made in both research and applied settings on topics that range from recruitment and selection (due to online platforms that provide new sources of data and advanced analytics to assess worker attributes for predicting performance) to diversity and inclusion (through multiobjective optimization techniques). They also detail the sources of data that organizational psychologists are integrating and using (e.g., enterprise data, social media data) and the various algorithms that are used to analyze those data (e.g., machine learning, text mining). This chapter showcases the extensive adoption of big data in the field of industrial and organizational psychology and the benefits it provides.

Part IV, Recommendations for Responsible and Rigorous Use of Big Data, addresses practical and ethical challenges associated with big data methods such as privacy, data security and storage, data sharing, and replicability and reproducibility issues. By providing information on technical advances and ethical considerations pertaining to big data, this provides guidance to psychologists undertaking big data research.

In Chapter 15, Paxton discusses key ethical principles of conducting human subject research as described in the Belmont Report, as well as their implications for today's psychological science dealing with big data. She introduces issues and open questions for readers to consider, which go beyond the scope of the Belmont Report (e.g., how to achieve a balance between requirements of open science and protection of participant rights, securing computational pipelines). In doing so, the chapter serves as a helpful guide for shaping the field's discussion about big data research ethics. We expect this work to be informative for updating our ethical principles in light of big data.

In Chapter 16, Strauss and Grand discuss key characteristics of robust and reliable science (e.g., relevant, rigorous, replicable and cumulative, transparent and open, theory oriented), providing a useful framework for evaluating big

data science in psychology. This is especially timely given that the fashionability of big data methods can lead to the relaxing of research standards when evaluating such work. They discuss extensions of the topics—HARKing, questionable research practices, and replicability and reproducibility—to big data, providing recommendations to enhance the rigor of big data science.

In Chapter 17, Bauer, Truxillo, Jones, and Brady address issues of privacy and cybersecurity as they relate to the specific context of big data application: personnel selection via online application systems. Considering both the benefits and concerns for multiple stakeholders in this context, the authors provide a set of practical recommendations for online job applicants, employers, developers and providers of online job application tools, and policymakers. They also highlight areas in which best-practice recommendations do not yet exist and thus need further research.

In Chapter 18, Bertino itemizes the numerous ways that the privacy of individuals can be compromised in the collection, storage, and analysis of big data. Her chapter is organized around privacy enhancing techniques that are also consistent with maintaining a high level of data security. She describes specific computer science techniques for addressing privacy and security concerns and identifies several remaining challenges in reconciling security with privacy. These challenges relate to data confidentiality (for which the emphasis is control of access) and data privacy (for which the major issue is that unforeseen information can be obtained about individuals and populations by correlating multiple big data sets). She also briefly discusses data transparency—a key requirement in today's big data era. Although much is being done to ensure the security and privacy of data, big data researchers have to understand that it is a complex problem.

In the closing chapter, Oswald summarizes key themes that emerge across the different chapters in the book. Although there is considerable promise in big data, particularly in allowing investigation of complex human behavior, there are also limitations that psychology can potentially help overcome. The limitations include the interpretability of data models, which can benefit from a substantive grounding in psychology; predictive biases that can profit from psychologists' engagement in open science practices that seek to balance transparency and privacy; and the privacy concerns themselves, which can be addressed through ethical and legal frameworks in psychology. Oswald also suggests multiple future research directions to evaluate the effectiveness of psychological science in improving the quality of big data projects (e.g., measurement, interpretation, research design, transparency) and data science teams (e.g., training and evaluation of teams).

To summarize, our goal in preparing this edited volume has been to provide frank discussions of the many potential benefits for psychological researchers opened up by big data, as well as possible pitfalls. As such, the book is intended for students (undergraduates and graduates), researchers, and practitioners who are interested in what the future of psychology holds in light of the "fourth industrial revolution" (i.e., digital technology). We hope that readers will come away with an appreciation for various ways in which they may be

able to incorporate big data research into their research programs and into the multidisciplinary research programs in which they engage.

REFERENCES

- Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, 73, 899–917. <http://dx.doi.org/10.1037/amp0000190>
- Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Retrieved from <https://www.wired.com/2008/06/pb-theory/>
- Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
- Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21, 447–457. <http://dx.doi.org/10.1037/met0000120>
- Hickman, L., Tay, L., & Woo, S. E. (2019). Validity investigation of off-the-shelf language-based personality assessment using video interviews: Convergent and discriminant relationships with self and observer ratings. *Personnel Assessment and Decisions*. Retrieved from https://www.researchgate.net/publication/332544721_Validity_Investigation_of_Off-the-Shelf_Language-Based_Personality_Assessment_using_Video_Interviews_Convergent_and_Discriminant_Relationships_with_Self_and_Observer_Ratings
- Woo, S. E., Tay, L., Hickman, L., & Saef, R. (2019). *More questions about the validity of machine-learning approaches to personality assessment: A case of social media text mining*. Manuscript submitted for publication.