Psychological Methods
2016, Vol. 21, No. 4, 447–457

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000120

Big Data in Psychology: Introduction to the Special Issue

Lisa L. Harlow
University of Rhode Island

Frederick L. Oswald
Rice University

The introduction to this special issue on psychological research involving big data summarizes the highlights of 10 articles that address a number of important and inspiring perspectives, issues, and applications. Four common themes that emerge in the articles with respect to psychological research conducted in the area of big data are mentioned, including: (a) The benefits of collaboration across disciplines, such as those in the social sciences, applied statistics, and computer science. Doing so assists in grounding big data research in sound theory and practice, as well as in affording effective data retrieval and analysis. (b) Availability of large data sets on Facebook, Twitter, and other social media sites that provide a psychological window into the attitudes and behaviors of a broad spectrum of the population. (c) Identifying, addressing, and being sensitive to ethical considerations when analyzing large data sets gained from public or private sources. (d) The unavoidable necessity of validating predictive models in big data by applying a model developed on 1 dataset to a separate set of data or hold-out sample. Translational abstracts that summarize the articles in very clear and understandable terms are included in Appendix A, and a glossary of terms relevant to big data research discussed in the articles is presented in Appendix B.

Keywords: big data, machine learning, statistical learning theory, social media data, digital footprint

Big data involves the storing, retrieval, and analysis of large amounts of information and has been gaining interest in the scientific literature writ large since the 1990s. As a catch-all term, big data has also been referred to by a number of other related terms such as: data mining, knowledge discovery in databases, data or predictive analytics, or data science. The domain has traditionally been associated with computer science, statistics, and business, and now it is clearly, quickly, and usefully making inroads into psychological research and applied practice. There is a healthy and growing infrastructure for dealing with big data, some of it being open source and free to use. For example, Hadoop (a name originally based on that of a child's toy elephant) is a widely used open source file system and framework. Within this framework, *MySQL* is a structured query language that is also open source and is used a great deal. It allows powerful capabilities to "select" a specific group of entities, "from" a specific database or set of files, "where" one or more specific conditions hold. For example, an academic researcher could select and analyze data based on student identification numbers from class records in several majors, where the GPA is less than 2.0. In turn, this could allow for the possibility of strategic data-driven interventions with these students to offer enrichment or tutoring that would bolster their grades and improve their chances of staying in school and succeeding. Once big data are queried and refined, they can be analyzed with a number of tools, increasingly with commonly known software and programs such as R and Python, respectively.

Who is using big data? Business industries in this area abound (e.g., insurance, manufacturing, retail, pharmaceuticals, transportation, utilities, law, gaming, eBay, telecommunication, hotels). Social media is also prominently involved (e.g., Google, Facebook, LinkedIn, Yahoo, Twitter). Various academic disciplines also have a visible presence (e.g., genomics, medicine, and environmental sciences, the latter often using spatial geographic information systems, or GIS). There are several journals in this area, including the open access and peer-reviewed journal *Big Data*, founded in 2013 and currently edited by Dhar. Their web page (http://www.liebertpub.com/overview/big-data/611/) boasts a comprehensive coverage and audience—yet has not yet mentioned psychology or even the broader social sciences. At least two other journals were founded in 2014, the open access *Journal of Big Data* that is edited by Furht and Khoshgoftaar, and *Big Data Research* that is edited by Wu and Palpanas. Likewise, these two journals also do not appear to be directed to those in psychology or the larger social sciences. Similarly, a quick Google search in September 2016 for "big data book" revealed more than 48 million results, although it is noteworthy that all of the big data books listed on the front page, are not specifically directed to social science fields. Noting all of this is not to indict the current state of

Lisa L. Harlow, Department of Psychology, University of Rhode Island; Frederick L. Oswald, Department of Psychology, Rice University.

A draft of a portion of this introduction was previously presented in Harlow, L. L., & Spahn, R. (2014, October). *Big data science: Is there a role for psychology?* Abstract for Society of Multivariate Experimental Psychology, Nashville, TN. The co-editors (Harlow and Oswald) would like to thank the authors and reviewers who contributed to this special issue. We also would like to offer much appreciation and thanks to our manuscript coordinator, Meleah Ladd, who has played an integral part in helping to make every aspect of our work better and more enjoyable, and especially so with this special issue. Lisa Harlow also extends thanks to the National Institutes of Health grant G20RR030883.

Correspondence concerning this article should be addressed to Lisa L. Harlow, Department of Psychology, University of Rhode Island, Kingston, RI 02881-0808. E-mail: lharlow@uri.edu

big data for neglecting psychology—quite the opposite: Psychology and the social sciences should be proactive and take advantage of a real opportunity in front of them. The timing is ripe, now that the big data movement has matured beyond many of its fads.

So, where does psychology fit into the field of big data or related areas such as computational social science? There are a number of areas in which psychology could and has begun to weigh in, such as wellness, mental health, depression, substance use, behavioral health, behavior change, social media, workplace well-being and effectiveness, student learning and adjustment, and behavioral genetics. A number of recent books of interest to psychology researchers have been published (Alvarez, 2016; Cioffi-Revilla, 2014; Matz Mayer-Schönberger & Cukier, 2013; McArdle & Ritschard, 2014, to name a few). Researchers are studying topics such as health and the human condition in big data sets comprising thousands of individuals, such as in the Kavli Human Project (http://kavlihumanproject.org/; Azmak et al., 2015). In a similar vein, Fawcett (2015) discusses the analysis of what is called the quantified self in which individuals collect data on themselves (e.g., number of steps, heart rate, sleep patterns) using personal trackers such as Fitbit, Jawbone, iPhone, and similar devices. Researchers envision studies that could link such personal data to health and productivity to reveal patterns or links between behavior and various outcomes of interest.

It is apparent that big data or data science is here to stay, with or without psychology. This broad-and-growing field offers a unique opportunity for interested psychological scientists to be involved in addressing the complex technical, substantive, and ethical challenges with regard to storing, retrieving, analyzing, and verifying large data sets. Big data science can be instrumental in collaboratively working to uncover and illuminate cogent and robust patterns in psychological data that directly or indirectly involve human behavior, cognition, and affect over time and within sociocultural systems. These psychological patterns, in turn, give meaning to nonpsychological data (e.g., medical data involving health-related interventions, booms and busts tied to financial investing behavior). The big data community, and big data themselves, can together propel psychological science forward.

In this special issue, we offer 10 articles that focus on various aspects of big data and how they can be used by applied researchers in psychology and other social science fields. One of the common themes of these articles is also clearly evident in federal funding announcements for big data projects: Psychologists and psychology benefit from the collaboration and contributions of other disciplines—and vice versa. For example, such collaborations can incorporate cutting-edge breakthroughs from computer science that can help access and analyze large amounts of data, as well as theory and behavioral science from across the social sciences that offer insight into the areas that are most in need of understanding, prediction, and intervention.

A second theme is that data are widely available in open forums such as Facebook, Twitter, and other social media sites, and can offer the opportunity to identify trends and patterns that are important to address. For example, tapping the content of Google activity could indicate geographic areas where users are inquiring about various flu or other symptoms, thus pointing to areas in which it may be important to focus health intervention efforts. The psychological nature of the query content might allow for early planning in targeting the intervention (e.g., judging the level of knowledge and concern about the health problem and its related symptoms and treatment). Note that when big data analyses incidentally detect a useful signal in the noise of social media data, one's discoveries and research efforts need not stop there; researchers can develop new construct-driven measures that help amplify those signals that may have initially been discovered serendipitously.

A third general theme is that it is critically important to consider and carefully attend to the ethical issues of big data projects, including data acquisition and security, the protection of the identity of the users who often inadvertently provide extensive data, and decisions about how the information will be used and interpreted vis-à-vis the nature of the audience or stakeholders involved.

A fourth shared theme of these articles is that it is essential to develop theories and hypotheses on an initial training set of data and then verify those findings with other validation data sets, either from a hold-out sample of the original data or from separate, independent data. With the existence of large data sets that often may not have had an overriding theory or set of hypotheses guiding their formation, an initial analysis of big data is often at the exploratory or data mining level. At least one or more subsequent analyses of separate data may be needed to be able to generalize past the initial data, particularly as there can be a large number of variables that are relevant to prediction, but not necessarily the best measures that one could obtain with additional foresight and planning. Given a large number of incidental variables, and given the flexible modeling afforded by big data analyses, it is perhaps more important than ever to avoid overinterpreting what might be considered a modern-day version of the classic "*crud factor*" (Meehl, 1990, p. 108), namely where researchers could find the appearance of relationships between variables in a large dataset that are robustly upheld (e.g., through cross-validation), yet these relationships may change or dissipate over time, as the nature of the relevant sample, population, and the phenomenon under study change as well. Each of the articles in this special issue address one or more of these four themes in relatively easy-to-understand presentations of how big data can be used by researchers in psychology.

A summary of the highlights of the articles is presented, below, followed by Appendix A, which provides translational abstracts (TAs) of the articles, briefly describing the essence of the articles in clearly understandable language. Appendix B includes a glossary of some of the major terms used in the 10 articles, providing brief descriptions of each and an indication of which articles refer to these terms. To be clear, the glossary is not intended to provide an exhaustive list of big data concepts; it is more of a summary of some of the ideas and practices that are referred to in these special issue articles so that readers can have a reference of the terminology and find out which special issue articles are discussing them. To help identify which terms are included in the glossary in Appendix B, these terms are italicized in this introductory article, although not necessarily in the separate articles themselves.

The first article by Chen and Wojcik (2016) offers an excellent guide to conducting behavioral science research on large data sets. In addition to describing some background and concepts, they provide three tutorials in the supplemental materials in which interested readers can move through the steps. Their first tutorial clearly indicates how to acquire the congressional speech data

through *application programming interfaces (APIs)* that reflect specific procedures needed to acquire data from a site. Their second tutorial demonstrates how these data are analyzed using procedures known as *latent semantic analysis (LSA)* and *latent Dirichlet allocation (LDA)* topic modeling, both of which can be used to assess the co-occurrence of words in a dataset based on underlying topics and relationships between documents. Other terms, common to the big data community and discussed in their main article and their third tutorial, include *bag of words*, *stop words*, *support vector machines*, *machine learning*, and s*upervised learning algorithms* (see also our glossary in Appendix B of this article). Chen and Wojcik also provide two appendixes to help apply the material they discuss. Their Appendix A provides the Python code for acquiring data from the Congressional Daily Digest that are discussed in the first and second tutorials, and the use of *MySQL*. Their Appendix B offers a checklist for conducting research with big data.

In the second article, Landers et al., (2016) discuss *web scraping*, an automated process that can quickly extract data from websites behind the scenes. Behavioral scientists are increasingly involved in this type of research, within academia and in organizations, determining the pulse of social consciousness and norms on web sources such as Facebook, Twitter, Instagram, and Google. Along with delineating potential benefits of web scraping, Landers et al. also provide their expert advice on the need to emphasize theory in such a project. In particular, they discuss what they call *theory of the data source* or *data source theory* to help ensure the relevance and meaningfulness of data that are obtained from web scraping. Although there are not yet exact standards on the ethics of scraping the web for data, Landers et al. suggest that the *APA Ethical Principles of Psychologists and Code of Ethics* (APA, 2010), along with those from the *Data Science Association*, can suggest policies and procedures for collecting data in a responsible manner that respects the participants and the research field in which conclusions will be shared. Assessing large data sets that are gleaned or scraped from the web using the theory-driven method suggested by Landers et al. can help lessen the possibility that the findings are just happenstances of a large collection of information.

The third article, by Kosinski et al., (2016), discusses how to use large databases collected from the web to understand and predict a relevant outcome. Their article is a tutorial that describes an example of using Facebook *digital footprint* data, stored in what is called a *user-footprint matrix*, to predict personality characteristics. The authors analyze input from over 100,000 Facebook users (see *myPersonality project,* http://www.mypersonality.org/; Kosinski, Matz, Gosling, Popov, & Stillwell, 2015) using dimension-reduction procedures such as *singular value decomposition (SVD)* that is computationally easy to use as a method for conducting *principal components analysis.* The Kosinski et al. article also discusses the LDA clustering procedure, also discussed in Chen and Wojcik to help form dimensions with similar content from large data sets of text or counts of words or products. Findings from an LDA model can be visually depicted in a *heatmap* that shows darker colors when a trait or characteristic is more correlated with one of the LDA clusters. Thus, you can see at glance the patterns that characterize each cluster.

In the fourth article, Kern et al., (2016) discuss the analysis of big data found on social media, such as on Facebook and Twitter. The authors discuss several steps in acquiring, processing, and quantifying these kinds of data, so as to make them more manageable for statistical analyses. The authors discuss a *World Well-Being Project* and use LDA and LSA (see also Chen & Wojcik) to reduce large amounts of text-based information into a smaller set of relevant dimensions. They also discuss a procedure known as *differential language analysis*, encouraging the use of *database management systems* that pervade the world of business and increasingly are being implemented in psychological research. Cautioning that results could be specific to a particular dataset and need to be further tested with independent data, Kern et al. explain and implement the *k-fold cross-validation* method that tests a prediction model across repeated subsets of a large dataset to support the robustness of the findings. The authors also discuss prediction methods such as the *lasso* (i.e., *least absolute shrinkage and selection operator*) as a regression method for robust prediction, based on screening a large set of predictors and weighting predictors that were selected conservatively (i.e., with lower magnitudes than traditional OLS regression). They also caution against *ecological fallacies*, whereby researchers derive erroneous conclusions about individuals and subgroups based on results from a larger group of data, and *exception fallacies*, when a conclusion is drawn based on outliers (exceptions) in the data that may stand out but may not fully represent the group. Not everyone uses social media, and some use it far more often or idiosyncratically than others. Still, these authors are optimistic about the amount and richness of the data that can be gleaned from social media, and the insights that can be gained from such data.

In the fifth article, Jones, Wojcik, Sweting, and Silver (2016) examine the content of Twitter posts after three different traumatic events (violence in or near college campuses), applying linguistic analyses to the text for negative emotional responses. They discuss a procedure known as *Linguistic Inquiry and Word Count* and an R-based computer *twitteR package* to analyze such data. Using an innovative approach, the authors recognize pertinent Twitter users by identifying people who follow relevant community networks tied to the geographical area of the event, and they are careful to compare results with control groups not similarly geographically situated, to help ensure that results were event-driven, versus other contemporaneous events that were more geographically widespread. Overall, this work demonstrates how psychological themes can be reliably extracted and related to region- and time-dependent events, similar to prior related work in the health arena.

In the sixth article, Stanley and Byrne (2016) contribute a theory-driven approach to big-data modeling of human memory (i.e., long-term knowledge storage and retrieval), testing two theoretical models that predict the tags that users apply to Twitter and *Stack Overflow* posts. Incorporating but going beyond the psychological tenet that "past behavior predicts future behavior," the current models robustly predict how and to what extent this tenet applies given the nature, recency, and frequency of past behavior. This article exemplifies an important general point, that big-data analyses benefit from being theory-driven, demonstrating how theories can develop in their usefulness as a joint function of empirical competition (i.e., deciding which model affords better prediction) and empirical cooperation (i.e., demonstrating how model ensembles might account for the data more robustly than

models taken individually). The authors discuss the use of an *ACT-R based Bayesian model* and a *random permutation model* to understand and clarify predictions about links between processes and outcomes.

The seventh article, by Brandmaier et al., (2016) discuss *ensemble methods* that they developed, one of which is called *structural equation model (SEM) trees* that combines *decision trees* (also called *recursive partitioning methods*) and SEM to understand the nature of a large dataset. These authors suggest an extended method called *SEM forests* that allows researchers to generate and test hypotheses, combining both data- and theory-based approaches. These and other methods, such as *latent class analysis* and *multiple sample SEM*, help in assessing distinct clusters in the data. Several methods are described to gauge how effectively an SEM forest is modeling the data, such as examining *variable importance* based on *out-of-bag* samples from the SEM trees, as well as *case proximity* and conversely, an *average dissimilarity* metric, the latter indicating its *novelty*. Brandmaier et al. provide two examples to demonstrate the use of SEM forests. Interested researchers can conduct similar analyses using Brandmaier's (2015) *semtree package* that is written in R, with their supplemental material providing the R code for the examples they provide.

In the eighth article, Miller, Lubke, McArtor, and Bergeman (2016) detail a new method for detecting robust nonlinearities and interactions in large data sets based on *decision trees*. Called *multivariate gradient boosted trees*, this method extends a well-established *machine-learning* or *statistical learning theory* method. Whereas most predictive models in the big data arena try to explain a single criterion, the present approach considers multiple criteria to be predicted (as does the Beaton et al. *partial least squares correspondence analysis* method). Such exploration is useful for informing and refining theories, measures, and models that take a more deductive approach. To do this, a boosted tree-based model for each outcome is fit separately, where the goal is to minimize cross-validated prediction error across all outcomes. An advantage of tree-based methods comes in detecting complex predictive relationships (interactions and nonlinearities) without having to specify their functional form beforehand. In the current approach, tree models can be compared across outcomes, and the explained covariance between pairs of outcomes can also be explored. The authors illustrate this approach using measures of psychological well-being as predictors of multiple psychological and physical health outcomes. Interested readers can apply this method to their own data with Miller's R-based *mvtboost package*.

In the ninth article, Chapman, Weiss, and Dubenstein (2016) consider measure-development models that focus squarely on predictive validity using a *machine-learning* approach that challenges—and complements—traditional approaches to measure development involving psychometric reliability. The proposed approach seeks out additional model complexity so long as it is justified by increased prediction; the approach incorporates *k-fold cross-validation* methods to avoid model overfitting. Almost two decades ago, McDonald's (1999) classic book, *Test Theory: A Unified Treatment*, also suggested that measures of a construct judged to be similar should not only demonstrate psychometric reliability, but also show similar relationships with measures of other constructs in a larger nomological net. The current big-data article reflects one important step toward advancing this general idea, discussing procedures and terms such as *elastic net, expected prediction error*, *generalized cross-validation error*, *stochastic gradient boosting* and *supervised principal components analysis,* as well as R-based computer packages *glmnet*, and *superpc*.

For the final tenth article, Beaton, Dunlop, and Abdi (2016) jointly analyze genetic, behavioral, and structural MRI in a tutorial for a generalized version of partial least squares called *partial least squares correspondence analysis* (PLSCA). The method can handle disparate data types that are on widely different scales, as might become increasingly common in large and complex data sets. In particular, their methods can accommodate categorical data when analyzing relationships between two sets of multivariate data, where traditional analyses assume the data for each variable are continuous (or even more strictly, multivariate normal). These authors have developed a freely available R package, *TExPosition*, which allows readers to apply the PLSCA method to their own data.

In closing, we hope you find something of interest to you in one or more of the 10 articles we present in this special issue on the use of big data in psychology. We recognize that other articles may approach these topics differently, and likewise, many other big data topics will be discussed in the future. We look forward to continued tutorials and other research publications in *Psychological Methods* that share even more about how to apply innovative and informative big data methods to meaningful and relevant data of interest to researchers in psychology and related social science fields.

References

Alvarez, R. M. (2016). *Computational social science: Discovery and prediction*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781316257340

APA. (2010). *Ethical principles of psychologists and code of conduct*. Retrieved from http://www.apa.org/ethics/code/

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., & Patrinos, A. (2015). Using big data to understand the human condition: The Kavli HUMAN Project. *Big Data, 3,* 173–188. http://dx.doi.org/10.1089/big.2015.0012

Bair, E., & Tibshirani, R. (2010). *superpc: Supervised principal components. R package version 1.07*. Retrieved from http://www-stat.standord.edu/~tibs/superpc

Beaton, D., Dunlop, J., & Abdi, H. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods, 21,* 621–651.

Brandmaier, A. M. (2015). *semtree: Recursive partitioning of structural equation models in R* [Computer software manual]. Retrieved from http://www.brandmaier.de/semtree

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21,* 566–582.

Chapman, B. P., Weiss, A., & Duberstein, P. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods, 21,* 603–620.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods, 21,* 458–474.

Cioffi-Revilla, C. (2014). *Introduction to computational social science: Principles and applications*. London, UK: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4471-5661-1

Fawcett, T. (2015). Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data, 3,* 249–266. http://dx.doi.org/10.1089/big.2015.0049

Friedman, J., Hastie, T., Simon, N., & Tibshirani, R. (2016). *glmnet: Lasso and elastic-net regularized generalized linear models. R package 2.0–6*. Retrieved from https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Gentry, J. (2016). *Package "twitteR," version 1.1.9*. Retrieved from https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

Jones, N. M., Wojcik, S. P., Sweeting, J., & Silver, R. C. (2016). Tweeting negative emotion: An investigation of Twitter data in the aftermath of violence on college campuses. *Psychological Methods, 21,* 526–541.

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods, 21,* 507–525.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70,* 543–56. http://dx.doi.org/10.1037/a0039210

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

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods, 21,* 475–492.

Matz Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston, MA: Houghton Mifflin Harcourt.

McArdle, J. J., & Ritschard, G. (Eds.). (2014). *Contemporary issues in exploratory data mining in the behavioral sciences*. New York, NY: Routledge.

McDonald, R. P. (1999). *Test theory: A unified treatment*. New York, NY: Routledge.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141. http://dx.doi.org/10.1207/s15327965pli0102_1

Miller, P. J., Lubke, G. H., McArtor, D. B., & Bergeman, C. S. (2016). Finding structure in data using multivariate tree boosting. *Psychological Methods, 21,* 583–602.

Stanley, C., & Byrne, M. D. (2016). Comparing vector-based and Bayesian memory models using large-scale datasets: User-generated hashtag and tag prediction on Twitter and Stack Overflow. *Psychological Methods, 21,* 542–565.

Appendix A

Translational Abstracts (TAs) for the 10 Special Issue Articles

1. TA for "A Practical Guide to Big Data Research in Psychology" by Eric Evan Chen and Sean P. Wojcik

The massive volume of data that now covers a wide variety of human behaviors offers researchers in psychology an unprecedented opportunity to conduct innovative theory- and data-driven field research. This article is a practical guide to conducting big data research, covering the practices of acquiring, managing, processing, and analyzing data. It is accompanied by three tutorials that walk through the acquisition of real text data, the analysis of that text data, and the use of an algorithm to classify data into different categories. Big data practitioners in academia, industry, and the community have built a comprehensive base of tools and knowledge that makes big data research accessible to researchers in a broad range of fields. However, big data research does require knowledge of software programming and a different analytical mindset. For those willing to acquire the requisite skills, innovative analyses of unexpected or previously untapped data sources can offer fresh ways to develop, test, and extend theories. When conducted with care and respect, big data research can become an essential complement to traditional research.

2. TA for "A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data from the Internet for Use in Psychological Research" by Richard N. Landers, Robert C. Brusso, Katelyn J. Cavanaugh, and Andrew B. Collmus

One of the biggest challenges for psychology researchers is finding high quality sources of data to address research questions of interest. Often, researchers rely on simply giving surveys to undergraduate students, which can cause problems when trying to draw conclusions about human behavior in general. To work around these problems, sometimes researchers actually watch people in real life, or observe via the web, taking notes on their behaviors to be analyzed later. But this process is time-consuming, difficult, and error-prone. In this article, we provide a tutorial on a technique that can be used to create data sets summarizing actual human behavior on the Internet in an automated way, partially solving both of these problems. This big data technique, called web scraping, takes advantage of a programming language called Python commonly used by data scientists. We also introduce a new related concept, called data source theories, as a way to address a common criticism of many big data approaches—specifically, that because the analytic techniques are "data-driven," they tend to take advantage of luck more so than psychology's typical approaches. As a result of this tendency, researchers sometimes make conclusions that do not reflect reality beyond their dataset. In creating a data source theory, researchers precisely account why the data they found exist and test the hypotheses implied by that theory with additional analyses. Thus, we combine the strengths of psychology (i.e., high quality measurement and rich theory) with those of data science (i.e., flexibility and power in analysis).

(*Appendices continue*)

3. TA for "Mining Big Data to Extract Patterns and Predict Real-Life Outcomes" by Michal Kosinski, Yilun Wang, Himabindu Lakkaraju, and Jure Leskovec

Humans are increasingly migrating to the digital environment, producing large amounts of digital footprints of behaviors, communication, and social interactions. Analyzing big data sets of such footprints presents unique methodological challenges, but could greatly further our understanding of individuals, groups, and societies. This tutorial provides an accessible introduction to the crucial methods used in big data analysis. We start by listing potential data sources, and explain how to efficiently store and prepare data for the analysis. We then show the reader how to reduce the dimensionality and extract patterns from big data sets. Finally, we demonstrate how to employ such data to build prediction models. The text is accompanied by examples of R code and a sample dataset, allowing the reader put their new skills into practice.

4. TA for "Gaining Insights from Social Media Language: Methodologies and Challenges" by Margaret L. Kern, Gregory Park, Johannes C. Eichstaedt, H. Andrew Schwartz, Maarten Sap, Laura K. Smith, and Lyle H. Ungar

Many people spend considerable time on social media sites such as Facebook and Twitter, expressing thoughts, emotions, behaviors, and more. The massive data that are available provide researchers with opportunities to study people within their real-world contexts, at a scale previously impossible for psychological research. However, typical psychological methods are inadequate for dealing with the size and messiness of such data. Modern computational linguistics strategies offer tools and techniques, and numerous resources are available, but there is little guidance for psychologists on where to even begin. We provide an introduction to help guide such research. We first consider how to acquire social media data and transform it from meaningless characters into words, phrases, and topics. Both top-down theory driven approaches and bottom up data-driven approaches can be used to describe characteristics of individuals, groups, and communities, and to predict other outcomes. We then provide several examples from our own work, looking at personality and well-being. However, the power and potential of social media language data also brings responsibility. We highlight challenges and issues that need to be considered, including how data are accessed, processed, analyzed, and interpreted, and ever-evolving ethical issues. Social media has become a valuable part of social life, and there is much we can learn by cautiously bringing together the tools of computer science with the theories and insights of psychology.

5. TA for "Tweeting Negative Emotion: An Investigation of Twitter Data in the Aftermath of Violence on College Campuses" by Nickolas M. Jones, Sean P. Wojcik, Josiah Sweeting, and Roxane Cohen Silver

Capturing a snapshot of emotional responses of a community soon after a collective trauma (e.g., school shooting) is difficult. However, because of its rapid distribution and widespread use, social media such as Twitter may provide an immediate window into a community's emotional response. Nonetheless, locating Twitter users living in communities that have experienced collective traumas is challenging. Prior researchers have either used the extremely small number of geo-tagged tweets (3%–6%) to identify residents of affected communities or used hashtags to collect tweets without certainty of the users' location. We offer an alternative: identify a subset of local community Twitter accounts (e.g., city hall), identify followers of those accounts, and download their tweets for content analysis. Across three case studies of college campus killings (i.e., UC-Santa Barbara, Northern Arizona State University, Umpqua Community College), we demonstrate the utility of this method for rapidly investigating negative emotion expression among likely community members. Using rigorous longitudinal quasi-experimental designs, we randomly selected Twitter users from each impacted community and matched control communities to compare patterns of negative emotion expression in users' tweets. Despite variation in the severity of violence across cases, similar patterns of increased negative emotion expression were visible in tweets posted by followers of Twitter accounts in affected communities after the killings compared to before the violence. Tweets from community-based Twitter followers in matched control communities showed no change in negative emotion expression over time. Using localized Twitter data offers promise in studying community-level response in the immediate aftermath of collective traumas.

6. TA for "Comparing Vector-Based and Bayesian Memory Models Using Large-Scale Datasets: User-Generated Hashtag and Tag Prediction on Twitter and Stack Overflow" by Clayton Stanley and Michael D. Byrne

The growth of social media and user-created content on online sites provides unique opportunities to study models of long-term memory. By framing the task of choosing a hashtag for a tweet and tagging a post on *Stack Overflow* as a long-term memory retrieval problem, two long-term memory models were tested on millions of posts and tweets and evaluated on how accurately they predict a user's chosen tags. An uncompressed and compressed model (in terms of storage of information in long-term memory) were tested on the large data sets. The results show that past user behavior of tag use is a strong predictor of future behavior. Furthermore, past behavior was successfully incorporated into the compressed model that previously used only context. Also, an attentional weight term in the uncompressed model was linked to a natural language processing method used to attenuate common words (e.g., articles and prepositions). Word order was not found to be a strong predictor of tag use, and the compressed model performed comparably to the uncompressed model without including word order. This shows that the strength of the compressed model is not in the ability to represent word order, but rather in the way in which information is efficiently compressed. The results of the large-

(*Appendices continue*)

scale exploration show how the architecture of the two memory models can be modified to significantly improve accuracy, and may suggest task-independent general modifications that can help improve model fit to human data in a much wider range of domains.

7. TA for "Theory-Guided Exploration with Structural Equation Model Forests" by Andreas M. Brandmaier, John J. Prindle, John J. McArdle, and Ulman Lindenberger

Building models fully informed by theory is impossible when data sets are large and their relations to theory not yet specified. In such instances, researchers may start with a core model guided by theory, and then face the problem which additional variables should be included and which may be omitted. Structural equation model (SEM) trees, a combination of SEM and decision trees, offer a principled solution to this selection problem. SEM trees hierarchically split empirical data into homogeneous groups sharing similar data patterns by recursively selecting optimal predictors of these differences from a potentially large set of candidates. SEM forests are an extension of SEM trees, consisting of ensembles of SEM trees each built on a random sample of the original data. By aggregating the predictive information contained in a forest, researchers obtain a measure of variable importance that is more robust than corresponding measures from single trees. Variable importance informs on what variables may be missing from their models and may guide revisions of the underlying theory. In summary, SEM trees and forests serve as a data-driven tool for the improvement of theory-guided latent variable models. By combining the flexibility of SEM as a generic modeling technique with the potential of trees and forests to account for diverse and interactive predictors, SEM trees and forests serve as a powerful tool for hypothesis generation and theory development.

8. TA for "Finding Structure in Data Using Multivariate Tree Boosting" by Patrick J. Miller, Gitta H. Lubke, Daniel B. McArtor, and C. S. Bergeman

Collecting data from smart-phones, watches, or web sites is a promising development for psychological research. However, exploring these data sets can be challenging because there are often extremely large numbers of possible variables that could be used to predict an outcome of interest. In addition, there is often not much established theory that could help making a selection. Using statistical models such as regression models for data exploration can be inconvenient because these standard methods are not designed to handle large data. In the worst case, using simple statistical models can be misleading. For example, simply testing the correlation between predictors and outcomes will likely miss predictors with effects that are not approximately linear. In this article we suggest using a machine learning method called "gradient boosted decision trees." This approach can detect predictors with many different kinds of effects, but is easy to use compared with fitting many different statistical models. We extend this method to multivariate outcomes, and implement our approach in the R package *mvtboost* which is freely available on CRAN. To illustrate the approach, we analyze predictors of psychological well-being and show how to estimate, tune, and interpret the results. The analysis showed, for example, that especially above average control of internal states is associated with increased personal growth. Experimental results from statistical simulations verify that our approach identifies predictors with nonlinear effects and achieves high prediction accuracy. It exceeds or matches the performance of other cutting edge machine learning methods over a wide range of conditions.

9. TA for "Statistical Learning Theory for High Dimensional Prediction: Application to Criterion-Keyed Scale Development" by Benjamin P. Chapman, Alexander Weiss, and Paul Duberstein

Researchers are often faced with problems that involve predicting an important outcome, based on a large number of factors that may be plausibly related to that outcome. Traditional methods for null hypothesis significance tests of one or a small number of specific predictors are not optimal for such problems. Machine learning, reformulated in a statistics framework as statistical learning theory (SLT), offers a powerful alternative. We review the fundamental tenets of SLT, which center around constructing models that maximize predictive accuracy. Importantly, these models prioritize predictive accuracy in new data, external to the sample used to build the models. We illustrate three common SLT algorithms exemplifying this principle, in the psychometric task of developing a personality scale to predict future mortality. We conclude by reviewing some of the diverse contexts in which SLT models might be useful. These contexts are unified by research problems that do not seek to test a single or small number of null hypotheses, but instead involve accurate prediction of an outcome based on a large amount of potentially relevant data.

10. TA for "Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data" by Darek Beaton, Joseph Dunlop, and Hervé Abdi

For nearly a century, detecting the genetic contributions to cognitive and behavioral phenomena has been a core interest for psychological research, and that interest is even stronger now. Today, the collection of genetic data is both simple and inexpensive. As a consequence a vast amount of genetic data is collected across different disciplines as diverse as experimental and clinical psychology, cognitive sciences, and neurosciences. However, such an explosion in data collection can make data analyses very difficult. This difficultly is especially relevant when we wish to identify relationships within, and between genetic data and, for example, cognitive and neuropsychological batteries. To alleviate such problems, we have developed a multivariate approach to

(*Appendices continue*)

make these types of analyses easier and to better identify the relationships between multiple genetic markers and multiple behavioral or cognitive phenomena. Our approach—called partial least squares correspondence analysis (PLSCA)—generalizes partial least squares and identifies the information common to two different data tables measured on the same participants. PLSCA is specifically tailored for the analysis of complex data that may exist in a variety of measurement scales (e.g., categorical, ordinal, interval, or ratio scales). In our article, we present—in a tutorial format—how PLSCA works, how to use it, and how to interpret its results. We illustrate PLSCA with genetic, behavioral, and neuroimaging data from the Alzheimer's disease Neuroimaging Initiative. Finally, we make available R code and data examples so that those interested can easily learn and use the technique.

Appendix B

Glossary of Some of the Major Terms Used in the 10 Special Issue Articles

ACT-R based Bayesian models are based on the ACT-R theory of declarative memory that can be operationalized as a big data predictive model, reflecting how declarative memory processes (e.g., exposure, learning, recall, forgetting) affect behavioral outcomes. The predictive model incorporates a version of the Naïve Bayes method, such that any piece of knowledge is assigned a prior probability for being retrieved by the user, independent of all other pieces of available knowledge, which is then weighted by the information in the current context to yield a posterior distribution and prediction (see Stanley and Byrne).

APA Ethical Principles of Psychologists and Code of Ethics (APA, 2010), along with those from the *Data Science Association*, suggest policies and procedures for collecting data in a responsible manner that respects the participants and the research field in which conclusions will be shared (see Landers et al., 2016).

Application Programming Interfaces (APIs) refer to sets of procedures that software programs use to request and access data in a systematic way from other software sources (APIs can be web based or platform-specific; see Chen and Wojcik 2016; Jones et al., 2016; Kern et al., 2016; and Stanley and Byrne 2016).

Average dissimilarity is a general term indicating how different a case tends to be from the rest of the data (see Brandmaier et al., 2016).

Bag of words conveys word frequency in a relevant text (e.g., sentence, paragraph, entire document), without retaining the ordering or context of the words (see Chen and Wojcik 2016).

Case proximity is a general term for the similarity between entities in a data set, identifying any clear outliers (see Brandmaier et al., 2016).

Crud factor (Meehl, 1990, p. 108) is a general term used to indicate that in any psychological domain, measures of constructs are all correlated with one another, at some overall level. Traditional analyses have dealt with this, as will big data analyses (see Harlow and Oswald).

Data source theory refers to a well-thought out theoretical rationale, developed on the basis of the available variables in a given set, to support the nature of the data and the findings derived from them. Researchers working with big data projects are encouraged to have a data source theory to guide exploration, analyses, and empirical results in large data sets (see Landers et al., 2016).

Database management system (DBMS) is a structure that can store, update, and retrieve large amounts of data that can be accrued in research studies (see Kern et al., 2016).

Data Science Association (http://www.datascienceassn.org/) is an educational group that offers guidelines for researchers to follow regarding ethics and other matters relevant to organizations (see Landers et al., 2016).

Decision trees (also called recursive partitioning methods) are models that apply a series of cutoffs on predictor variables, such that at each stage of selecting a predictor and cutoff point, the two groups created by the cutoff are as separated (i.e., internally coherent and externally distinct) as possible on the outcome variable. Decision trees model complex interactions, because each split of the tree on a given predictor is dependent on all splits from the previous predictors (see Brandmaier et al., 2016, and Miller et al., 2016).

Differential language analysis (DLA) is an empirical method used to extract underlying dimensions of words or phrases without making a priori assumptions about the structure of the language, and then relating these dimensions to outcomes of interest (see Kern et al., 2016).

Digital footprint refers to data that can be obtained from various sources such as the web, the media, and other forums in which publicly available information is posted by or stored regarding individuals or events. These kind of data can be stored in what is called a *User-Footprint Matrix* (see Kosinski et al., 2016).

(*Appendices continue*)

Ecological fallacies are incorrect conclusions made about individual people or entities that are derived from information that summarizes a larger group. For example, if a census found that higher educational levels were associated with higher income, it would not necessarily be true that everyone with high income had a high level of education. Simpson's paradox is an extreme example, where each within-group relationship may be different from or even the opposite of a between-groups relationship (see Kern et al., 2016).

Elastic net refers to a regression model that linearly weights the penalty functions from two regression models: the *lasso* regression model (applying an L1 penalty that conducts variable selection and shrinkage of nonzero weights) and the ridge regression model (applying an L2 penalty that applies shrinkage, does not select variables, and will include correlated predictors, unlike *lasso*; see Chapman et al., 2016).

Ensemble methods involve the use of predictions across several models. The idea is that combining predictions across models tends to be an improvement over the predictions taken from any single model in isolation. An example of an ensemble method is the *structural equation model random forests* (see this term, below; see Brandmaier et al., 2016).

Exception fallacies involve mistaken conclusions about a group derived from a few unrepresentative instances in which an event, term, or characteristic occurs quite a lot. For example, if one or two participants in a dataset mention the word "sad" many times, it could falsely be surmised that the group of data as a whole experienced depression (see Kern et al., 2016).

Expected prediction error (EPE) is an index of accuracy for a predictive model, decomposed into: (a) squared bias (systematic model over- or underprediction across data sets); (b) variance (fluctuation in the model parameter estimates across data sets); and (c) irreducible error variance (variance that cannot be explained by any model). Expected prediction error captures the *bias-variance tradeoff*: Models that are too simple will underfit the data and show high bias yet low variance in the EPE formula; models that are too complex will overfit the data and show low bias yet high variance in the EPE formula (see Chapman et al., 2016).

Generalized cross-validation error indicates the target that is be minimized (the loss function) in *k*-fold cross-validation: for example, the sum of squared errors, the sum of absolute errors, or the Gini coefficient for dichotomous outcomes (see Chapman et al., 2016).

glmnet is a computer package written in R code by Friedman, Hastie, Simon, and Tibshirani (2016) that fits *lasso* and *elastic-net* models, with the ability to graph model solutions across the entire path of relevant tuning parameters (see Chapman et al., 2016).

Heatmaps plot the relationships among variables and/or clusters, using colors or shading to indicate the strength of relationship among variables (see Kosinski et al., 2016).

k-fold cross-validation involves partitioning a large dataset into *k* subsets of equal size. First, a model is developed on (*k*-1) partitions of the data—the "test" data set; then predicted values from model are obtained on the *k*th partition of the data that was held out—the "training" data set). This process is repeated *k* times so that all the data serve as training data, and all data therefore have predicted values from models in which they did not participate (see Chapman et al., 2016, and Kern et al., 2016).

Lasso (least absolute shrinkage and selection operator) is a regression method that helps screen out predictor variables that are not contributing much to a model relative to the others (see Kern et al., 2016).

Latent class analysis can help explain the heterogeneity in a set of data by clustering individuals into unobserved types, based on observed multivariate features. Features may be continuous or categorical in nature (see Brandmaier et al., 2016).

Latent Dirichlet allocation (LDA) is a method that models words within a corpus as being attributable to a smaller set of unobserved categories (topics) that are empirically derived (see Chen and Wojcik 2016).

Latent semantic analysis (LSA) involves the examination of different texts, where it is assumed that the use of similar words can reveal common themes across different sources (see Chen and Wojcik 2016, Kosinski et al., 2016 and Kern et al., 2016)

Linguistic inquiry and word count (LIWC) is a commercial analysis tool for matching target words (words within the corpus being analyzed) to dictionary words (words in the LIWC dictionary). Target words are then characterized by the coded features of their matching dictionary words, such as their tense and part of speech, psychological characteristics (e.g., affect, motivation, cognition), and type of concern (e.g., work, home, religion, money; see Jones et al., 2016).

Machine learning, which has also been called *statistical learning theory,* is a generic term that refers to computational procedures for identifying patterns and developing models that improve the prediction of an outcome of interest (see Chapman et al., 2016; Chen & Wojcik 2016; Harlow & Oswald; Kern et al., 2016; and Miller et al., 2016).

Multiple sample structural equation modeling (SEM) helps in testing differences across the different clusters that emerge, to identify the patterns of heterogeneity (see Brandmaier et al., 2016).

Multivariate gradient boosted trees involve a nonparametric regression method that applies the idea of stochastic gradient boosting to trees (see *stochastic gradient boosting*). Trees are fitted iteratively to the residuals obtained from previous trees, while seeking to optimize cross-validated prediction across multiple outcomes (not just one; see Miller et al., 2016).

(*Appendices continue*)

mvtboost is a package written in R code by Miller that implements multivariate gradient boosted trees, allowing the user to tune and explore the model (see Miller et al., 2016).

MyPersonality project (http://www.mypersonality.org/; Kosinski, Matz, Gosling, Popov, & Stillwell, 2015) stores the scores from dozens of psychological questionnaires as well as Facebook profile data of over six million participants (see Kosinski, et al., 2016).

MySQL is an open source version of a structured query language for working with big data projects (see Harlow and Oswald, and Chen and Wojcik 2016).

Novelty refers to how different a case is from the rest of the data, showing little proximity and more dissimilarity (see Brandmaier et al., 2016).

Out-of-bag samples are portions of a larger dataset that do participate in the development of a predictive model and can be used to generate predicted values (and error). Out-of-bag samples are similar to the test sample data referred to previously in *k*-fold cross-validation (see Brandmaier et al., 2016).

Partial least squares correspondence analysis (PLSCA) is a generalization of partial least squares that can extract relationships from two separate sets of data measured on the same sample. In particular, PLSCA is useful for handling both categorical and continuous data types (e.g., genetic single-nucleotide polymorphisms that are categorical, and behavioral data that are roughly continuous). Permutation tests and bootstrapping are applied to conduct statistical inference for the overall fit of the model as well as inference on the stability of each obtained component (see Beaton et al., 2016).

Random permutation model is an approach for determining whether to preserve information about word order in text analytics, in case doing so provides additional predictive information. Permutations create uncorrelated vectors as a point of contrast with the actual ordering (see Stanley and Byrne 2016).

semtree is a computer package that was developed by (Brandmaier, 2015; http://brandmaier.de/semtree/) and written in R. It can be used to analyze SEM tree and forest methods to help explore and discern clusters or subgroups within a large dataset (see Brandmaier et al., 2016 and related references).

Singular value decomposition (SVD) is a procedure used to reduce a large set of variables or items to a smaller set of dimensions. It is one approach to conducting a *principal components analysis* (see Kosinski et al., 2016).

Stack Overflow is an online question-and-answer forum for programmers (using R, Python, and otherwise; see Stanley and Byrne 2016).

Stochastic gradient boosting is a general term for an iterative method of regression, such that the predictor entered first has the highest functional relationship with the outcome; then residuals are created, and the same rule is applied (where the outcome now becomes the residuals). Also, at each iteration, only a subset of the data is used to help develop more robust models (where *out-of-bag* prediction errors can be obtained from the data outside of the model). The learning rate and number of iterations are, loosely, inversely related (low learning rate, or improvement in prediction at each step, generally means more iterations) and optimizing these can be explored and supported through cross-validation (see Chapman et al., 2016).

Stop words are words that are not essential to a phrase or text and therefore can be omitted to help keep a file more concise. Examples of stop words include "an" and "the" or other similarly nondescript words that can be deleted from a large database (e.g., Twitter, Facebook) and do not need to be analyzed (see Chen and Wojcik 2016; Kern et al., 2016; and Stanley and Byrne 2016).

Structural equation model (SEM) forests are classification procedures that combine SEM and decision-tree or SEM-tree methods to understand the nature of subgroups that exist in a large dataset. SEM forests extends the method of SEM trees by resampling the data to form aggregates of the SEM trees that should have less bias and more stability (see Brandmaier et al., 2016).

Structural equation model (SEM) trees combine the methods of decision trees and SEM to conduct theory-guided analysis of large data sets. SEM trees are useful in examining a theoretically based prediction model, but can be unstable when random variation in the data is inadvertently featured in a decision tree (see Brandmaier et al., 2016).

superpc is a computer package written in R code by Bair and Tibshirani (2010) that conducts the procedure known as supervised principal components analysis, a term that is defined below (see Chapman et al., 2016).

Supervised learning algorithms are procedures that can be developed on a training dataset, and then be used to build regression models that can predict an outcome with one or more variables (see Chen and Wojcik 2016).

Supervised principal components analysis (SPCA) is a generalization of principal components regression that first selects predictors with meaningful univariate relationships with the outcome and then performs principal components analysis. Cross-validation is used to determine the appropriate threshold for variable selection and the number of principal components to retain (see Chapman et al., 2016).

(*Appendices continue*)

TExPosition is a computer package written in R code by Beaton and colleagues that implements *partial least squares correspondence analysis* (this latter term being defined previously; see Beaton et al., 2016).

Theory of the data source is the process whereby a larger conceptual framework is adopted when analyzing and interpreting findings from a large dataset, particularly one obtained for another purpose, such as with web scraping of generally available data (see Landers et al., 2016).

twitteR is a package written in R code by Jeff Gentry (2016) that accesses the Twitter API (see glossary entry on this term), which then allows one to extract subsets of Twitter data found online, search the data, and subject the data to text analyses (see Jones et al., 2016).

User-footprint matrix holds information obtained from sources such as the web or various records and lists (see Kosinski et al., 2016).

Variable importance is a term indicating how much the inclusion of a specific variable will reduce the degree of uncertainty there is in a model (or models) of interest. The uncertainty criterion and the model of course must be mathematically formalized (see Brandmaier et al., 2016).

Web scraping is a process that culls large amounts of data from web pages to be used in observational or archival data collection projects (see Landers et al., 2016).

World Well-Being Project (WWBP, http://www.wwbp.org/) involves a collaboration with researchers from psychology and computer science. The project draws on language data from social media to study evidence for well-being that can be revealed through themes of interpersonal relationships, successful achievements, involvement with activities, and indication of meaning and purpose in life (see Kern et al., 2016).

Received October 4, 2016
Accepted October 5, 2016 ■