Editorial

Challenges of Big Data Analyses and Applications in Psychology

Mike W.-L. Cheung[1] and Suzanne Jak[2]

[1] Department of Psychology, Faculty of Arts and Social Sciences, National University of Singapore, Singapore
[2] Methods and Statistics, Child Development and Education, University of Amsterdam, The Netherlands

Because psychological research commonly involves gathering data from human subjects, psychological studies traditionally do not involve (very) large samples. The sample sizes of most experimental studies are relatively small (median sample size around 40; Bakker, Van Dijk, & Wicherts, 2012), whereas the sample sizes of observational studies tend to be small to medium (median sample size around 120; Fraley & Marks, 2007). Various statistical procedures have been developed to make inferences based on data with small to medium sample sizes. However, it may be expected that more and more big data or large datasets will be available in psychology in the future. Big data may be available in the form of experimental or observational data. For experimental studies, researchers may, for example, use tracking devices to frequently capture participants' physiological and psychological responses over a certain period. A large amount of observational data may be available when researchers extract data from Websites such as Facebook, Twitter, Google, and Amazon. Many governments and organizations, for example, the EU Open Data Portal (http://data.europa.eu/euodp/en/home) and the US government (https://www.data.gov/), have also made their data freely accessible to the public. Another area in which the volumes of data are rapidly getting larger is human brain imaging, where developments in MRI hardware and software provide possibilities to increase the spatial resolution of images, the number of images, as well as the number and duration of scanning sessions, leading to massive amounts of data (Smith & Nichols, 2018). In general, we may expect that more and more big data will be available to psychologists shortly, leading to an increasing need to reflect on the methodological and statistical issues related to the analysis of big data.

Big data raise questions about the quality of the data and the generalizability of results. In research where data are gathered through the Internet or other media, researchers may have little control over who is providing the data, and information about the background characteristics of participants may be difficult to obtain. This will lead to uncertainty about the population for which the research findings will apply (Hargittai, 2015). Moreover, with big sample sizes, it is a concern that even a small selection bias may lead to a false rejection of the null hypothesis (Ioannidis, 2005).

Besides reflecting on data quality, big data call for the development of suitable statistical tools. Conventional statistical methods taught in graduate schools (e.g., Aiken, West, & Millsap, 2008) may not be sufficient in the era of big data. Several issues can be involved when applying conventional statistical methods to big data. First, the datasets may be too large for standard workstations, so that researchers' computers cannot be used to handle the data. Second, the data may involve different types of information such as numerical data, text, sound, and videos. Conventional multivariate techniques such as multiple regression may not be optimal for handling such different types of variables. Third, the large sample sizes may lead to statistically significant results in most statistical analyses, even when the associated effect sizes are practically trivial. If researchers use the significance test as the criterion in, for example, judging the relevance of the predictors, misleading conclusions can be made.

Machine learning techniques (e.g., Hastie, Tibshirani, & Friedman, 2008; Murphy, 2012) such as regression trees, regularized regression, random forests, neural networks, and support vector machines are popular in big data analytics. In contrast to conventional statistics (e.g., multiple regression), these techniques focus on the accuracy of prediction rather than the accuracy of the parameter estimates. There are two basic types of techniques – supervised and unsupervised learning. Supervised learning (e.g., regression trees and regularized regression) is used when the "correct" outcomes are known. Conceptually, computer programs are trained in predicting the correct outcomes by minimizing errors in future cases. Unsupervised learning (e.g., cluster analysis) is used to classify objects with similar features

https://doi.org/10.1027/2151-2604/a000348

empirically. The division between conventional multivariate statistics and machine learning techniques has been characterized as "the two cultures" in the statistics literature (Breiman, 2001; Donoho, 2017; Shmueli, 2010). Given the usefulness of machine learning techniques in making predictions, we may expect that more psychologists are going to apply some form of machine learning techniques in their research.

The collection of papers in this topical issue of *Zeitschrift für Psychologie* demonstrates how machine learning techniques can be used to address psychological research questions in the big data era. de Schipper and Van Deun (2018) propose a novel principal component analysis that takes data from different blocks into account. Big data may include information from different types or blocks of data. Examples include the genome-wide data or data from the Internet such as tweets, web page visits, and GPS signals. It is challenging to analyze and combine these data sources. By using two real examples and a simulation study, the authors show how the proposed method could be used to combine information to form components based on the predefined blocks of variables.

Schoedel et al. (2018) demonstrate how a combination of smartphone sensing data with traditional self-report data could be used to address research in personality. Specifically, they used the behaviorally focused log data in participants' mobile phones and self-reported data to predict participants' sensation seeking. The authors found that the smartphone logging data could predict the self-reported sensation seeking scores.

Pargent and Albert-von der Gönna (2018) show how to apply machine learning techniques to address psychological research questions using a large dataset from the German GESIS Panel (Bosnjak et al., 2018). They utilized several machine learning techniques to predict panelists' gender, sick days, evaluation of US President Donald Trump, income, life satisfaction, and sleep satisfaction. The authors also provide details of various critical choices in applying machine learning techniques. These include, for example, the choice of preprocessing data, model classes, resampling techniques, hyperparameter tuning, and performance measures.

Using the same GESIS Panel, Vijayakumar and Cheung (2018) asked a slightly different research question. They attempted to address how replicable the findings of machine learning are. Specifically, they tested the replicability of the variable selection, along with the predictive accuracy, of some machine learning techniques including support vector machines, random forests, multivariate adaptive regression splines, regularized regression variants (least absolute shrinkage and selection operator (LASSO) and elastic net), and multiple regression using an empirical dataset and simulated data. They found that the machine learning techniques performed better than multiple regression when the structures were complicated with nonlinear patterns whereas multiple regression performed well in selecting predictors with a weak association with the criterion.

While the above papers focus on how machine learning techniques can be applied to psychological research, Zhang, Liu, Xu, Yang, and Zhang (2018) apply the split/analyze/meta-analyze (SAM) approach developed by Cheung and Jak (2016) to handle big data in psychological research. The SAM approach breaks up a large data pool into smaller datasets that can be handled by conventional multivariate techniques on standard desktop computers. The results of each small dataset are then combined meta-analytically. The authors illustrate how the SAM approach could be used to predict the financial performance of companies by using company and country level data.

The papers in this topical issue highlight the potential of applying machine learning techniques to big data in psychology. There are still many pressing issues in this field. For example, there is a need to succinctly summarize the current state of applications of big data in psychology. Also needed are more tutorial-style papers (e.g., Chen & Wojcik, 2016; Landers, Brusso, Cavanaugh, & Collmus, 2016) on how to acquire and analyze big data from popular sources such as Facebook, Twitter, and Amazon Mechanical Turk (MTurk). Another important direction for future research is a comparison of the advantages and disadvantages of conventional multivariate statistics and machine learning techniques for big data in psychology. We hope that the papers presented in this issue will stimulate more research in big data in psychology.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*, 32–50. https://doi.org/10.1037/0003-066X.63.1.32

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543–554. https://doi.org/10.1177/1745691612459060

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review, 36*, 103–115. https://doi.org/10.1177/0894439317697949

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*, 199–215. https://doi.org/10.1214/ss/1009213725

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods, 21*, 458–474. https://doi.org/10.1037/met0000111

Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology, 7*, 738. https://doi.org/10.3389/fpsyg.2016.00738

de Schipper, N. C., & Van Deun, K. (2018). Revealing the joint mechanisms in traditional data linked with big data. *Zeitschrift für Psychologie, 226*, 212–231. https://doi.org/10.1027/2151-2604/a000341

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*, 745–766. https://doi.org/https://doi.org/10.1080/10618600.2017.1384734

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 149–169). New York, NY: Guilford Press.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science, 659*, 63–76. https://doi.org/10.1177/0002716215570866

Hastie, T., Tibshirani, R. J., & Friedman, J. (2008). *The elements of statistical learning* (2nd ed.). New York, NY: Springer.

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

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. https://doi.org/10.1037/met0000081

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Pargent, F., & Albert-von der Gönna, J. (2018). Predictive modeling with psychological panel data. *Zeitschrift für Psychologie, 226*, 246–258. https://doi.org/10.1027/2151-2604/a000343

Schoedel, R., Au, Q., Völkel, S. T., Lehmann, F., Becker, D., Bühner, M., . . . Stachl, C. (2018). Digital footprints of sensation seeking: A traditional concept in the big data era. *Zeitschrift für Psychologie, 226*, 232–245. https://doi.org/10.1027/2151-2604/a000342

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

Smith, S. M., & Nichols, T. E. (2018). Statistical challenges in "big data" human neuroimaging. *Neuron, 97*, 263–268. https://doi.org/10.1016/j.neuron.2017.12.018.

Vijayakumar, R., & Cheung, M. W.-L. (2018). Replicability of machine learning models in the social sciences: A case study in variable selection. *Zeitschrift für Psychologie, 226*, 259–273. https://doi.org/10.1027/2151-2604/a000344

Zhang, Y. E., Liu, S., Xu, S., Yang, M. M., & Zhang, J. (2018). Integrating the split/analyze/meta-analyze (SAM) approach and a multilevel framework to advance big data research in psychology: Guidelines and an empirical illustration via the human resource management investment–firm performance relationship. *Zeitschrift für Psychologie, 226*, 274–283. https://doi.org/10.1027/2151-2604/a000345

History
Published online February 22, 2019

Acknowledgments
We thank Edgar Erdfelder (the Editor-in-Chief) and Christina Sarembe (the Editorial Production Manager) for their advice and support in the preparation process. We also acknowledge Michael Bosnjak (Director of ZPID – Leibniz Institute for Psychology Information and University of Trier) for organizing the associated conference on "Big Data in Psychology 2018" that took place June 7–9, 2018 in Trier, Germany.

Funding
Mike W.-L. Cheung was supported by the Academic Research Fund Tier 1 (FY2017-FRC1-008) from the Ministry of Education, Singapore. Suzanne Jak was supported by the Netherlands Organization for Scientific Research (NWO-VENI-451-16-001).

Mike W.-L. Cheung
Department of Psychology
Faculty of Arts and Social Sciences
National University of Singapore
Block AS4, Level 2, 9 Arts Link
Singapore 117570
mikewlcheung@nus.edu.sg

Suzanne Jak
Methods and Statistics
Child Development and Education
University of Amsterdam
Nieuwe Achtergracht 127
1018 WS Amsterdam
The Netherlands
s.jak@uva.nl