Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Big data methods in the social sciences

Frederick L Oswald[1] and Dan J Putka[2]

Addresses
[1] Rice University, 6100 Main St – MS25, Houston, TX 77063, USA
[2] Human Resources Research Organization, 66 Canal Center Plaza #600, Alexandria, VA 22314, USA

Corresponding author: Oswald, Frederick L (foswald@gmail.com)

Current Opinion in Behavioral Sciences 2017, **18**:103–106

This review comes from a themed issue on **Big data in the behavioural sciences**

Edited by **Michal Kosinski** and **Tara Behrend**

For a complete overview see the Issue and the Editorial

Available online 21st November 2017

https://doi.org/10.1016/j.cobeha.2017.10.006

Big data methods are attractive to the social sciences — and all the sciences — because they can address data sets where the number of variables can far exceed the number of cases being analyzed. Generally speaking, big data methods seek to detect stable and potentially complex clusters and/or predictions in the data, while also taking aggressive steps not to capitalize on chance in doing so. On this latter point, big data methods often perturb some combination of the sample, the variables, and the model to ensure that solutions are robust, where traditional statistical methods rarely do so, with some exceptions (e.g., cross-validation approaches, latent variable models).

The boom in methodological advances and tools for big data affords the social scientist a vast and fast-evolving toolbox available for use. The goal of this annotated review is to provide some guidance to social science researchers who might feel overwhelmed by the panoply of big data methods available and/or feel that these methods are opaque. Despite feeling overwhelmed (or perhaps because of it), social scientists increasingly seek to understand big data methods and analyses that are sensible and appropriate to their research questions and data. This annotated review hopes to help this cause, relaying some key resources that

(a) build a fundamental understanding of key modeling methods available for big data;
(b) equip the researcher with a set of tools, based on these methods, that is available for conducting analyses of big data;
(c) provide an instructive sampling of relevant substantive applications of big data methodology in the social sciences.

The resources below were carefully selected, hoping to save readers valuable time as they engage in big data methodologies. Descriptions of each resource do not pretend to be comprehensive; instead, they serve as signals that encourage further exploration in a selective and strategic manner.

Conceptual introduction

Below, we refer to big data being analyzed using *supervised* or *predictive* models, where there is a criterion (or criteria) to be predicted. For instance, a health psychologist might seek to predict well being from patients' extensive medical and clinical intake data. Or an organizational psychologist might seek to predict employees' job performance measured over time, based on available HR data at the employee, team, and department levels. Examples of supervised methods include random forests and gradient boosted trees; neural networks and deep learning; and support vector machines.

Below, we also refer to the analysis of big data using *unsupervised* or *clustering* models, where the goal is to find structure within the data — the clusters are not self-evident but rather determined from the data, and further, there is no outcome to predict. These models would be relevant to a developmental psychologist who is seeking to find clusters of similar content both within and across personal diaries. As another example, a consumer psychologist might decide to cluster both movies and moviegoers on the basis of the movies' psychological content. Examples of unsupervised methods include principal components analysis (PCA), *k*-means clustering, and hierarchical partitioning.

References

Efron, B., & Hastie, T. (2016). *Computer age statistical inference* (Vol. 5). Cambridge University Press.

Authors are applied statisticians who are experts in the field, guiding the reader with accessible language through the history of statistics, first by walking the reader through both frequentist and Bayesian modes of statistical inference, then chronologically merging computational approaches to statistics that began in the 1950s. This useful foundation informs the final section of the book on predictive big data methods. Examples are supplied at the end of each chapter, many of them using R.

●● Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*, 78–87.

This 2012 paper is considered a classic by many, as a highly readable paper for both researchers and practitioners facing analytic problems with big data. The paper highlights some of the basic considerations (issues, questions, and pitfalls) typically faced, including (a) modeling, evaluation, and optimization choices; (b) the need for knowledge (not just data) to inform the development and use of big data models; (c) the curse of dimensionality (i. e., the more dimensions added, the less that even big data can populate the multidimensional space); (d) simpler algorithms based on more data tend to beat cleverer algorithms based on less data; and a final tenet that all social scientists learn by heart, that (e) correlation does not equal causation. We consider this paper an essential read for social scientists to gain their bearings around the topic of big data.

Technical resources

Social scientists have a wide range of free software available for analyzing big data. Those just learning about big data methods tend to use the R programming language because of the community of social science users and the relevant statistical packages available (https://www.r-project.org). The RStudio interface is often used to work in R (https://www.rstudio.com). Social scientists with more of a programming background may also use Python and its libraries for big data analysis, such as Pandas (https://pandas.pydata.org) and SciPy (https://www.scipy.org). We suspect that the Python community will grow in the social sciences, because it generalizes to other programming languages and other research communities, although we tend to focus on R here given the current state of affairs.

References

● Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer. Also see https://www.appliedpredictivemodeling.com.

This book provides a very practical and problem-oriented approach to applying and comparing different big data predictive models using the *caret* package in R. Authors help the reader through the data preparation process that is necessary prior to any data analysis (e.g., dealing with missing data, centering, and scaling). They then cover a variety of supervised and unsupervised methods. Case studies are offered throughout that, although not entirely social-science based, are quite valuable for exposing the reader to a step-by-step decision-making process for big data modeling.

●● James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer. Also see http://www-bcf.usc.edu/~gareth/ISL/.

Now in its seventh printing, this book is available as a PDF free for download at the website above, and it describes and applies a wide range of big data predictive models that are relevant across disciplines, such as finance, marketing, sociology, medicine, and genetics. The authors first provide a brief justification for the use of big data methods in terms of exploring complex relationships while retaining robustness via cross-validation approaches. Then the authors deliver a brief introduction to using R to allow the reader to carry out the examples of big data models that are both supervised and unsupervised in nature. This book is an indispensable resource, useful on its own or prior to using the Kuhn book mentioned previously, which focuses on supervised models.

●● Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, visualize, and model data*. Sebastopol, CA: O'Reilly Media. Also see http://r4ds.had.co.nz/.

As the title implies, this book makes use of R in tackling big data problems, but it goes beyond the mere execution of big data models by delving into how R can be applied to other critical efforts required prior to the analysis (data importing, management, and manipulation) as well as after the analysis (extracting, visualizing, and communicating one's results). This book is helpful as a single resource for appreciating and engaging in the full process of analyzing big data using R; also, it is available online at the website above, with its sections broken down into separate webpages. The lead author, Hadley Wickham, is an expert who not only created or co-created many of the popular R packages covered in the book (e.g., *tidy* for data management, *ggplot2* for graphing); he has spent thousands of hours in blogs, tutorials, and workshops that help others, and this experience comes through with this very clear and accessible resource.

●● Saltz, J. S., & Stanton, J. M. (2018). *An introduction to data science*. Los Angeles: SAGE.

This book provides an essential introduction for the social science researcher who is (like all of us once were) starting at ground zero, with no experience in using R and RStudio (a popular interface for R) or in applying big data methodologies. The author takes a clear and even humorous approach to using R; cleaning, managing, and manipulating data; applying big data methodologies; and using visualization and interactive applications. The book is replete with substantive examples (e.g., US Census data, Twitter data, consumer data), and big data methods that are both unsupervised (e.g., text mining, association mining), and supervised (e.g., support vector machines),