Behavior Research Methods (2019) 51:1473–1476
https://doi.org/10.3758/s13428-019-01278-2

EDITORIAL

Introduction to special issue. Beyond the lab: Using big data to discover principles of cognition

Gary Lupyan [1] · Robert L. Goldstone [2]

Published online: 8 July 2019

Like many other scientific disciplines, psychological science has felt the impact of the big-data revolution. This impact arises from the meeting of three forces: data availability, data heterogeneity, and data analyzability. In terms of data *availability*, consider that for decades, researchers relied on the Brown Corpus of about one million words (Kučera & Francis, 1969). Modern resources, in contrast, are larger by six orders of magnitude (e.g., Google's 1T corpus) and are available in a growing number of languages. About 240 billion photos have been uploaded to Facebook,[1] and Instagram receives over 100 million new photos each day.[2] The large-scale digitization of these data has made it possible in principle to analyze and aggregate these resources on a previously unimagined scale. *Heterogeneity* refers to the availability of different *types* of data. For example, recent progress in automatic image recognition is owed not just to improvements in algorithms and hardware, but arguably more to the ability to merge large collections of images with linguistic labels (produced by crowdsourced human taggers) that serve as training data to the algorithms. Making use of heterogeneous data sources often depends on their standardization. For example, the ability to combine demographic and grammatical data about thousands of languages led to the finding that languages spoken by more people have simpler morphologies (Lupyan & Dale, 2010). The ability to combine these data types would have been substantially more difficult without the existence of standardized language and country codes that could be used to merge the different data sources. Finally, *analyzability* must be ensured, for without appropriate tools to process and analyze different types of data, the "data" are merely bytes.

Our main goal in assembling this special issue was to highlight work that makes new types of data available or more accessible, that demonstrates creative merging of different types of data, and that describes new techniques that make it possible to draw useful inferences from the data, with a focus on advancing psychological theory. The call for contributions to this volume was broad, mentioning mining data from online databases and other "naturally occurring datasets" (Goldstone & Lupyan, 2016), the construction of new linguistic corpora, methods for analyzing diverse data sources, and the creation of environments for gamifying data collection (i.e., by making experiments fun, subject enrollment can be increased at no additional cost; Salganik, 2017). We deliberately omitted any restrictions on the number of subjects, stimuli, or observations, lest the concept of "big data" be reduced to "large files" (E.J. Ward, personal communication). We favored contributions that emphasized the impact of the new data sources/techniques on advancing psychological theory rather than viewing big data as an end in itself. Finally, we tried to maximize the usefulness of any new datasets and analytic techniques, by emphasizing the importance of open data, open experimental materials, and open code for analysis. Below we describe the contributions to this special issue, grouped into three broad clusters: uses of naturalistic or crowdsourced data, methodological advances (with an emphasis on improving data analyzability), and the creation of new data resources. We were pleased that many contributions could be placed in more than one cluster.

Uses of naturalistic and crowdsourced data

Several of the present articles test theory-driven predictions with naturalistic datasets. **Vinson, Dale, and Jones**

[1] https://mybilliondollarapp.com/how-many-photos-on-facebook-how-about-240-billion/

[2] https://www.omnicoreagency.com/instagram-statistics/

✉ Gary Lupyan
lupyan@wisc.edu

[1] Department of Psychology, University of Wisconsin, Madison, WI, USA

[2] Department of Psychological and Brain Sciences, Program in Cognitive Science, Indiana University, Bloomington, IN, USA

investigated decision contamination—how a person's decision influences later decisions—using millions of Yelp and Amazon reviews. **Kim et al.** used more than 10,000 employee training records to test for evidence of the spacing effect. **Cohen and Todd** examined dog adoption records, matching adopters' stated preferences with adoption decisions as a way to assess the stated versus revealed preference gap in the context of pet adoption; Rozin (2006) would be pleased. **Kumar et al.** tested several predictions from theories of autobiographical memories by applying text analysis to participants' emails and using sentences from emails from varying temporal intervals as retrieval cues in a memory experiment. An intriguing aspect of this work is the method by which the researchers accessed participants' Gmail accounts without compromising their privacy.

Steyvers and Benjamin analyzed 54 million plays of Lumosity "brain-training" games to examine effects of age and initial skill on performance and continued participation. The sheer density of the data allowed the authors to compare fits to alternative quantitative models in a way that would not be possible with typical lab-based studies. Also examining the effects of age on performance, **Vaci et al.** relied on the excellent player statistics kept by the National Basketball Association, finding that players who improve more also deteriorate more slowly, and that improvement is more variable than deterioration. The type of densely sampled longitudinal data they used would be extremely difficult and costly to obtain from lab studies. Vaci et al.'s study is an excellent demonstration of how psychologists can leverage the efforts of industries to collect and maintain high-quality data. Rather than complain to ourselves that in many countries sports matter more than science, we can take advantage of the enormous resources that have been poured into sports.

Several articles make use of existing text corpora as testing grounds for both general and specific predictions. **Hemmatian et al.** examined discourse on same-sex marriage between 2006 and 2017, a time period of rapid change in public opinion on the subject. By analyzing over 600,000 posts on Reddit (the self-proclaimed "front page of the Internet") and algorithmically classifying them as containing consequentialist versus protected-values discourse, the authors tried to answer the question of what kinds of discourse shifts accompany this rapid change in public opinion. In another creative use of Reddit, **Thorstad and Wolff** trained classifiers on language from discussion channels (subreddits) dedicated to different types of mental illness, showing that the classifiers quite accurately predict what channel a post came from (replicating past work), achieving above-chance performance even when classifiers were trained on posts from nonclinical subreddits. Moreover, the model trained on language from nonclinical posts was able to predict with moderate accuracy what clinical subreddit a user would post to later. Another example of using corpus data is **Johns and Dye**'s examination of gender biases in books and movies. Using corpora totaling over 2.5 billion words, the authors examine the distribution of male and female personal names, finding across almost all genres a strong bias for using male names. In a creative example of data aggregation, **Nalabandian and Ireland** test predictions from Narrative Arc Theory by applying automatic analysis to film scripts from the Internet Movie Script Database and relating them to movie reviews from Rotten Tomatoes and the Internet Movie Database. **Sagi** uses naturalistic language data (about 4000 19th century English books from Project Gutenberg) to test the natural partition hypothesis, stating that that context influences verb meanings more than noun meanings, and to test for the existence of phonesthemes. The results of the corpus study of phonesthemes are then used to design a lab experiment that tests whether people are sensitive to the relationships present in the corpus. This is an excellent example of "scaling-down" from naturalistic analyses to lab-based studies rather than the more commonly encountered attempts to scale up lab-based studies to the real-world; a type of neo-Gibsonian ecological psychology in a world where words make up much of our ecology.

Several articles report more traditional studies, using crowdsourcing to answer questions that would be far more laborious and expensive to answer using in-lab studies. **Cuskley et al.** recruited over 1,000 Dutch speakers—including over 200 synesthetes—to complete a vowel-to-color association task. The authors found systematic associations between colors and vowels and that these associations vary continuously from highly structured in synesthetes to individuals with no systematic mappings at all. These findings deserve to be singled out for the wonderful way in which they are visualized in the article and in an online interactive supplement. **Cipora et al.** used crowdsourcing to collect data from over 1,500 participants tested on varieties of SNARC/MARC effects. Although these effects have been well studied in the lab, the large samples available through crowdsourcing allowed the authors to examine interindividual differences—do people who map numbers to space in one way also map them in another way when different materials are used? The methods in Cuskley et al. and Cipora et al. are excellent examples of how at-scale recruitment made possible through crowdsourcing can drive systematic research of individual differences.

Although crowdsourcing has proven its worth time and again, we still occasionally hear critiques about how participants recruited through crowdsourcing such as Mechanical Turk are unrepresentative, make it impossible to verify that they understood the task, or could be completing the experiment in a highly uncontrolled environment. It is true that online samples are not representative, but they are far more representative in most ways than the typical convenience samples of freshman and sophomores currently enrolled in psychology

classes. Conducting studies online means giving up some of the control offered by a lab (other forms of control, such as accurate recording of response times and timed delivery of images and audio, can be accomplished with precision adequate for many purposes over web browsers). But giving up some control has its advantages, such as helping establish constraints on generality. If it turns out that the spacing effect can be demonstrated using messy naturalistic data, or that the SNARC effect can be measured precisely over a web browser in a sample that includes participants who are doing the task with screaming children in the background, so much the better!

We are reminded of Craik and Tulving's (1975) original investigation of depth of processing. After eight experiments using a rigorously controlled lab setting, "One of the authors, by nature more skeptical than the other, had formed a growing suspicion that this rigor reflected superstitious behavior rather than essential features of the paradigm. Accordingly, a simplified version of Experiment 2 was formulated which violated many of the rules observed in previous studies" (p. 287). The authors then detail the many ways in which traditional laboratory controls—individual testing, timed responses, controlled rate of stimulus presentation—were discarded. "The point of this study was not to attack experimental rigor, but rather to determine to what extent the now familiar pattern of results would emerge under the much looser conditions" (p. 287). Their findings were unchanged. Lab studies that use well-controlled manipulations are often uniquely powerful for inferring causation and distinguishing between competing mechanisms. But the public could be forgiven for not caring much about psychological phenomena that are obtained only under very specific conditions. Showing that a phenomenon can be obtained outside the lab is important for understanding its robustness. Discovering, often serendipitously, that factors that have been predicted to matter in fact do not—or (more likely) the reverse—is important for advancing theory.

Methodological advances

A number of articles in this issue describe new tools or analytic techniques and offer workflow guidance to researchers wishing to make use of big data. **Snefjella et al.** show how text embeddings trained on language from different time periods can be used to estimate time-specific lexical norms (in their case, word concreteness). The authors applied this technique to test claims that English has been becoming more concrete over the last 150 years. **Hsu et al.** describe a Markov chain Monte Carlo procedure with human agents to derive mental representations of categories spanning facial expressions, visual representations of the seasons, and dimensions of morality. This is precisely the kind of procedure that benefits from the ability to easily recruit large numbers of participants on well-specified tasks. **Solomon et al.** show how applying network-analytic methods to (crowd-sourced) concept features can illuminate the nature of conceptual representations. Among their results is a finding that the density of local feature associations within a concept network reduces the extent to which word meaning can vary across instances. **Frey et al.** show how affect (measured through sentiment analysis) spreads in the course of real-time communication. They do this by analyzing hundreds of thousands of timestamped messages exchanged on a social gaming platform. The proprietary nature of the data and the fact that the participants were 8- to 12-year-old children raise a number of potential privacy concerns that are addressed at length by the authors.

Van Dam and De Palma describe several projects that involve applying analysis of child-directed speech (recorded using a LENA system) using automated analyses and sharing the data with the public. The article includes a discussion of the trade-offs between coarse data from many people versus higher-resolution data from a few. As with Frey et al.'s contribution, their collection, processing, and sharing of children's data required resolving additional privacy considerations (see also the contributions, below, by Dennis et al.).

Three contributions describe new applications or frameworks for facilitating large-scale data collection. **Andreotta et al.** describe a workflow for qualitative text analysis of social media data. The number of questions that can be asked through analyses of social media data is equaled by the feeling of being overwhelmed by the amount and diversity of the data. Andreotta et al.'s systematic workflow promises to help. **Hartshorne et al.** evaluate the pros and cons of different methods for collecting large-scale online data and describe Pushkin, a new, highly extensible framework for running large-scale studies in browsers (including on mobile devices). The platform provides numerous mechanisms for recruiting, engaging, and retaining research subjects and provides methods for engaging the public in conducting citizen science. **Coris et al.** present RECAPP-XPR—a smartphone application for studying episodic memory over longer durations (days or weeks) than are typically manipulated in the lab.

Data creation, aggregation, and curation

Several contributions focus on the creation of new data or aggregating existing data in novel ways, or providing new interfaces for accessing and aggregating existing data. **Dennis et al.** present unforgettable.me, a platform that seeks to aggregate people's data from hundreds of sources, such as email, fitness trackers, GPS, and social media profiles. Understandably, privacy and data ownership are major concerns. The authors sketch a solution in which researchers create data queries, with participants retaining the rights to all of

their data. An especially intriguing part of this solution is automatic data censoring if the requested data end up being specific enough to personally identify the subject (see also **Dennis et al.**'s other contribution, "Privacy Versus Open Science," for a more extended discussion of privacy issues). **Buchanan et al.** present an extended collection of feature norms, totaling the number of concepts for which detailed English norms are available to over 4,400. **Li et al.** present The Macroscope, a tool that helps visualize and analyze historical changes in language by taking advantage of the enormous collections of digitized text that are now available to researchers.

To be useful, data not only have to exist, but they have to be easy to find and accessible. **Buchanan et al.**'s second contribution helps researchers navigate the increasing number of language norms that are available, by describing the Linguistic Annotated Bibliography, an annotated and searchable database of norms and stimuli. **Laarmann-Quante et al.** present the Litkey Corpus, a longitudinal corpus of written language produced by 8- to 11-year-old children. The rich annotations, flexible interface for querying the corpus, and openness of the data provide welcome changes from other corpora of children's writing (e.g., the Oxford Children's Corpus; Banerji, Gupta, Kilgarriff, & Tugwell, 2013) that remain locked down. Given the ubiquity of spoken language in our environment, it is remarkable how few of these data are in a format that is amenable to empirical study. **MacWhinney** describes the current state of TalkBank, a repository of spoken language (which includes the well-known child-language repository, CHILDES), containing large amounts of both auditory recordings and transcribed data for both adult and child-directed speech. Until recently, accessing TalkBank data required using a proprietary system or scraping the raw data. **Sanchez et al.**'s article describes childes-db, an R package for accessing CHILDES data in an R environment, thereby obviating the need to learn an additional system and increasing the usefulness of CHILDES.

Taken together, the contributions to this special issue of *Behavior Research Methods* provide updates and surprising extensions to theoretical developments in ecological psychology. Ecological psychology, as originally promulgated by Neisser (1976) and Gibson (1979), was premised on the need for psychology to study behavior in real-world, not only in laboratory environments. The contemporary extension of this call to study behavior in the real world, as exemplified by the articles in this issue, is to note that a large amount of a modern person's real world consists of language, people, and technological innovations. To get a complete understanding of human behavior, we need to understand human–environment interactions, where the environment crucially includes our cell phones, instant messages, online communities, email, movies, online games, sporting contests, and computers. As the present articles attest, all of these environmental components can provide a cornucopia of data when creatively and diligently investigated. The tools and analyses developed in these articles offer diverse perspectives that would be impossible to achieve in the laboratory: diverse measures taken over diverse conditions in diverse contexts from a diverse sample of participants. These multiple diversities will allow researchers of the future to much more effectively study representative samples of behavior. These samples will generalize to real-world behavior much more robustly than our traditional laboratory methods can, in no small part because they are taken from real-world behavior. As we stated earlier, we are not interested in equating "big data" with file sizes exceeding a particular threshold, but we do find the endeavor of "embiggening" behavioral science to include the kinds of diverse contexts, behaviors, tasks, and purposes found in the real world to be both worthwhile and exciting.

References

Banerji, N., Gupta, V., Kilgarriff, A., & Tugwell, D. (2013). Oxford Children's Corpus: A corpus of children's writing, reading, and education. In A. Hardie & R. Love (Eds.), *Corpus Linguistics 2013: Abstract book* (pp. 315–318). Lancaster, UK: (Lancaster) University Centre for Computer Corpus Research on Language. Retrieved from http://ucrel.lancs.ac.uk/cl2013/doc/CL2013-ABSTRACT-BOOK.pdf

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Goldstone, R. L., & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring data sets. *Topics in Cognitive Science*, *8*, 548–568. https://doi.org/10.1111/tops.12212

Kučera, H., & Francis, W. N. (1969). Computational analysis of present-day American English. *International Journal of American Linguistics*, *35*, 71–75.

Lupyan, G., & Dale, R. A. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*, e8559. https://doi.org/10.1371/journal.pone.0008559

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York, NY: W. H. Freeman/Times Books/Henry Holt & Co.

Rozin, P. (2006). Domain denigration and process preference in academic psychology. *Perspectives on Psychological Science*, *1*, 365–376. https://doi.org/10.1111/j.1745-6916.2006.00021.x

Salganik, M. J. (2017). Bit by bit. Retrieved June 20, 2019, from the Princeton University Press website: https://press.princeton.edu/titles/11057.html