


## Special Issue

---

# Validity Concerns in Research Using Organic Data

Heng Xu 

Nan Zhang

American University

Le Zhou

University of Minnesota

---

*With the advent of computing technologies, researchers across social science fields are using increasingly complex methods to collect, process, and analyze data in pursuit of scientific evidence. Given the complexity of research methods used, it is important to ensure that the research findings produced by a study are robust instead of being affected significantly by uncertainties associated with the design or implementation of the study. The field of metascience—the use of scientific methodology to study science itself—has examined various aspects of this robustness requirement for research that uses conventional designed studies (e.g., surveys, laboratory experiments) to collect data. Largely missing, however, are efforts to examine the robustness of empirical research using “organic data,” namely, data that are generated without any explicit research design elements and are continuously documented by digital devices (e.g., video captured by ubiquitous sensing devices; content and social interactions extracted from social networking sites, Twitter feeds, and click streams). Given the growing popularity of using organic data in management research, it is essential to understand issues concerning the usage and processing of organic data that may affect the robustness of research findings. This commentary first provides an overview of commonly present issues that threaten the validity of inferences*

---

*Acknowledgments: The authors are deeply grateful to Mo Wang and Gwendolyn K. Lee for their encouragement and guidance in the process of developing this manuscript. They are also grateful for helpful feedback and comments from the two anonymous reviewers as well as the participants of NSF Workshop on Promoting Robust and Reliable Research Practice. The authors would also like to thank Elizabeth M. Campbell and John D. Kammeyer-Mueller for their valuable comments and suggestions. Heng Xu’s and Nan Zhang’s work on this manuscript was supported in part by the National Science Foundation under Grant 1760059. Le Zhou’s work on this manuscript was supported in part by the National Science Foundation under Grant 1734134. Any opinions, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.*

*Corresponding author: Heng Xu, American University, 4400 Massachusetts Ave. NW, Washington, DC, 20016, USA*

*E-mail: xu@american.edu*

*drawn from empirical studies using organic data. This is followed by a discussion on some key considerations and suggestions for making organic data a robust and integral part of future research endeavors in management.*

**Keywords:** *open science (e.g., transparency in research practices); research methods; research design; replication studies*

---

With the advent of computing technologies, specifically the rapidly growing capability for data analytics, researchers across social science fields are using increasingly complex methods to collect, process, and analyze data in pursuit of scientific evidence. Given the complexity of research methods used, it is important to ensure that the research findings produced by a study are *robust* instead of being affected significantly by uncertainties associated with the design or implementation of the study. Various aspects of this robustness requirement, and various ways to address it, have been raised and debated in multiple disciplines closely related to management (e.g., Open Science Collaboration, 2015, in social psychology) under different terms, like *reproducibility*, *replicability*, *conceptual cross-validation*, *methodological triangulation*, and so on, which were often used interchangeably without a consensus on their definitions across different research fields (see Goodman, Fanelli, & Ioannidis, 2016). In this commentary, we use *robustness* to refer to the extent to which research findings (and inferences made by researchers) change as choices made in the study process vary.<sup>1</sup>

Despite the recent rise of awareness on issues related to robustness, social sciences in general, and the management field in particular, have paid limited attention to the robustness of findings from studies using “organic data.” While there is not yet a universally accepted, precise definition of the term *organic data* (as we shall elaborate later in this commentary), the common understanding in the research community is that the term refers to data *not* collected following an explicit research design but documented by a technology, device, or interface capturing natural “digital footprints” of human activities, such as data from sensing devices, mobile applications, or online social networks (Groves, 2011). So far, the number of studies using these digital footprints in management research remains relatively small. Nonetheless, it is important for researchers to recognize and anticipate potential issues associated with the usage of organic data so that practices for promoting robust research can be established ahead of the curve.

A variety of factors can lead to the lack of robustness in findings of a research study. For example, selecting statistical tests based on knowledge of test results and reporting only those tests returning statistically significant results obviously weaken the robustness of the research findings, as the findings have been manipulated by the researcher(s) and the method chosen is tied to the specific data set (Aguinis, Ramani, & Alabduljader, 2018; Nelson, Simmons, & Simonsohn, 2018). Instead of enumerating all possible concerns, our commentary focuses on one issue that is particularly pronounced for studies using organic data: the quality of inferences drawn in a study (i.e., the *validity* of the study). While there is a rapid emergence of studies using organic data, it is not always clear to researchers whether certain inferences are acceptable or, as called in metascience, invalid due to “questionable research practices.” Drawing from Shadish, Cook, and Campbell’s (2002) validity framework, there are four types

of validity that researchers are generally concerned about: statistical conclusion validity, internal validity, construct validity, and external validity. We will discuss the implications of practices that are common in the organic-data research process for the appropriateness, accuracy, and strength of support for inferences, particularly those inferences made about constructs (i.e., *construct validity*) and causal relationships (i.e., *internal validity*).

Outside of management research, concerns about threats to validity in the usage of organic data have already surfaced. In industrial-organizational psychology, researchers have noticed threats to construct validity from how organic data are processed in the measurement stage (Braun & Kuljanin, 2015). In political science and computer science, researchers have critiqued the practices of drawing inferences beyond the range that can be supported by the data at hand (King & Zeng, 2007) or from organic data with biased representation of certain subpopulations without proper adjustments in statistical analyses (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). Compared with such existing work, our goal is not to zoom in on one specific practice. Instead, we will highlight a broad spectrum of threats to validity that are often embedded in research using organic data, which are relatively rare in conventional designed experiments or surveys. These threats call for specific attention from management researchers starting to incorporate organic data into their research programs. We also propose potential solutions that are not yet commonplace in management research. In the rest of this commentary, we first briefly describe the key features of organic data and a typical workflow of studies using organic data. We then discuss potential validity threats and offer solutions.

## Overview of Validity Threats in Studies Using Organic Data

### *Definition of Organic Data*

As mentioned earlier, organic data are characterized by the lack of conventional research design elements during the data generation stage. Instead, organic data mostly document certain naturally occurring activities through technological devices or platforms (Groves, 2011). It may be tempting to simply treat organic data and conventional designed data as a dichotomy, but we believe multiple continuous dimensions better describe the various types of data used in management research (e.g., organic data, archival data, and data from laboratory experiments). These dimensions include (but are not limited to) the purpose of data generation (e.g., for a specific research study or for other purposes), the volume of data, existing data structures (i.e., how data are formatted), signal-to-noise ratio, and continuity in the assessment process. Taking this multidimensional view is important for at least two reasons.

First, certain forms of organic data and conventional designed data can be similar in some dimensions but differ in others. For example, U.S. Census data, corporate e-mail traces, and Twitter data are similar in the dimension of data generation purpose, as none of them is originally generated for a specific research study.<sup>2</sup> They are also similar on the dimension of data volume, as all can be voluminous. However, they differ on other dimensions, such as the existence of a clear data structure: Whereas all Census data and some e-mail traces are clearly structured (e.g., e-mails exchanged between mentors and newcomers during onboarding phase), Twitter data as well as some other e-mail traces (e.g., e-mails among a team of coworkers with fluid team membership over a long period of time) can be less structured and are best categorized as free text. Understanding the similarities and distinctions among various data types in different dimensions is important because it is these dimensions (or, more

precisely, features in these dimensions), rather than the type of data, that *drive research practices*. For example, the research practices of filtering input data with keywords or search patterns (e.g., regular expressions; see Aho, 1990) are often driven by the large volume of input data. As another example, the lack of structure in input data often drives researchers to adopt automated tools, such as information-extraction algorithms, to convert the input data into scores on given variables (Cafarella, Madhavan, & Halevy, 2009). Having a proper understanding of which features on which dimensions drive which research practices helps better identify common validity threats and makes the corresponding solutions applicable to multiple forms of data.

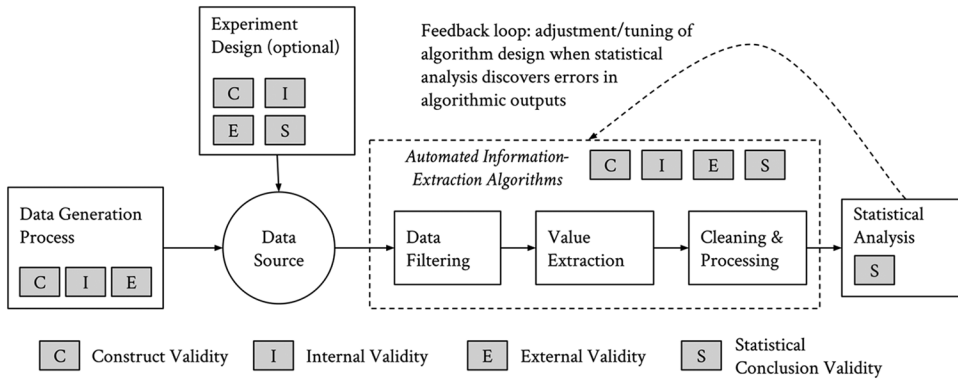
Second, our adoption of the term *organic data* coined by earlier researchers (e.g., Groves, 2011) is pragmatic, in order to raise awareness of an emerging trend in management research. As researchers start to conduct field experiments on online platforms (e.g., Facebook), either independently or in cooperation with platform providers, and attempt to establish causal relationships (e.g., Aral & Walker, 2011), the boundaries around “organic data” can become blurry. As we illustrate in the top part of Figure 1, in studies including experimentation on online platforms, data sets obtained for research reflect influences from both the manipulations used and other data generation mechanisms, making it less meaningful to categorize the obtained data as either “designed” or “organic.” Therefore, for observational and experimental studies alike, it is important to avoid attributing validity issues to the data being “organic” and instead understand the root issues (i.e., dimensions and features) that give rise to the validity threats.

### *Typical Workflow of Studies Using Organic Data and Threats to Validity*

The top section of Figure 1 depicts a typical workflow of studies using organic data. While most parts of the figure are self-explanatory, two parts differ from conventional experiments or surveys and may introduce validity threats not typically considered in conventional studies. The first is the data generation process, which is the “true model” underlying the observed pattern in the data (e.g., the “true” cause and effect between firm practices, employee personal attributes, and employee activities on social network sites; see discussions of the social media context in McFarland & Ployhart, 2015). The true model is never fully known by researchers, regardless of whether organic data are used. However, compared with conventional studies in a better-controlled environment (e.g., experiments in laboratories), the data generation process for organic data is even more opaque. This is because organic data are often affected by behaviors or decisions of the study participants or third parties (e.g., designers of websites) in addition to the researchers. Such behaviors and/or decisions are likely to be unknown to—and even unobservable by—the researchers. For example, users of wearable sensors (e.g., Fitbit) can manipulate or block data recording as study participants. Even more worrisome, large amounts of organic data may be deceptively created (Ott, Cardie, & Hancock, 2012), such as activities produced by automated software (i.e., “bots”) pretending to be human users.

The other important step that can differ from conventional studies is the use of automated tools to process the data, specifically the use of *automated information-extraction algorithms* (Cafarella et al., 2009) that convert input data (e.g., tweets from a given time period) into extracted “useful information” (as defined by the researcher; e.g., tweets mentioning certain events) and generate measurements that can be used for later statistical analyses (e.g., number of adjectives used). Researchers may use information-extraction algorithms designed

**Figure 1**  
**Typical Workflow of Studies Using Organic Data**



*Note.* Boxes represent typical steps and arrows represent general direction of workflow in studies using organic data, with the dashed arrow representing possible feedback loops in the research process. Circle represents the data source that studies operate on (e.g., e-mail traces, information recorded by social media platforms).

with a theory-driven approach to scrape data from web pages (Landers, Brusso, Cavanaugh, & Collmus, 2016), selectively download and filter the data based on the relevance to the research purpose, and produce values of study variables (e.g., levels of work stress) from the downloaded, often unstructured (e.g., data in a free-text format), data (see examples of applying automated information-extraction algorithms in Sajjadi, Sojourner, Kammeyer-Mueller, & Mykerezi, 2019; Wang, Hernandez, Newman, He, & Bian, 2016). Note in Figure 1 the feedback loop from statistical analysis back to the automated algorithms indicates the scenarios where the algorithm design has to be adjusted when further analysis of the extracted data unveils potential errors in the algorithmic outputs.

Threats to validity can arise in any part of the research process illustrated in Figure 1 (Shadish et al., 2002; Sussmann & Robertson, 1986). In the rest of this commentary, we focus on threats to validity that are associated with (a) the usage of automated information-extraction algorithms and (b) the organic data generation process, respectively. Note these two steps are orthogonal—organic data may be processed manually, while automated information-extraction algorithms (e.g., natural language processing [NLP], algorithms that provide automated analyses of text written in human languages) have been extensively used to process designed data, as well (e.g., Campion, Campion, Campion, & Reider, 2016). While the linkage between issues incurred by the opaqueness of the data generation process and different types of validity can be more easily identified and isolated, validity threats stemming from the adoption of automated information-extraction algorithms to replace manual efforts often cut across multiple types of validity (i.e., affecting multiple aspects of inferences drawn from the study). For example, when using NLP algorithms to process texts (e.g., responses to open-ended questions in job applications, performance reviews), the errors introduced by these algorithms can cause a variety of concerns. On the one hand, measurements produced by NLP algorithms often have poor accuracy when the text being processed contains irony or sarcasm, which affects construct validity (Davidov, Tsur, & Rappoport, 2010). On the other

hand, NLP algorithms have been found to produce lower error rates on data from some race/gender combinations than others (e.g., social media posts from minority and female vs. white and male; see Blodgett & O'Connor, 2017), threatening external validity, as well. As such, in the following, we start by discussing validity threats stemming from the usage of automated information-extraction algorithms, which cut across all aspects of validity. Then, we discuss validity threats incurred by the opaqueness of the data generation process, focusing on construct validity and internal validity.

## **Threats to Validity From the Usage of Automated Information-Extraction Algorithms**

We highlight two major threats in this category: errors in algorithmic outputs and a large parameter/procedure space of an information-extraction algorithm.

### *Threats From Errors in Algorithmic Outputs*

When “correct” answers can be identified based on certain standards (e.g., for the question of counting the number of words that include the letter *a*), if the outputs from an analytic system (e.g., human coders) deviate from the correct answer, we consider there is an error or a mistake. Just like human coders, automated algorithms designed to replace human coders in the information-extraction process can make mistakes. For some specific tasks, such as identifying the mentioning of an event from text data, like e-mails, the performance of state-of-the-art automated algorithms can rival that of untrained human coders (e.g., King & Lowe, 2003). However, for many other information-extraction tasks commonly used in management research, even the state-of-the-art automated algorithms today still produce significantly more errors than trained human coders (Schwaiger, Lang, Ritter, & Johannsen, 2016). While all errors may appear to be threats to validity, indeed, the use of numerous algorithms that incur a considerable amount of errors—far exceeding that of a human expert—poses no significant validity threat. The key determinant here is not the *amount* of errors but whether the nature of errors can be known to the researcher, who can then use such knowledge to account for the errors in design or statistical analyses and thereby minimize (or even eliminate) the validity threats incurred by the errors. In fact, a seemingly high accuracy rate of an algorithm could mask the potential threats and make them even more lethal.

*Benign errors.* Some errors in algorithmic outputs can be diagnosed by researchers and rectified through study design or statistical analyses. These can be considered benign errors. For example, consider a study on the relationship between team member composition and quality of software produced by a team of engineers. Instead of asking human experts to examine millions of lines of code to measure the quality of team outputs, a more practical solution is to use algorithms to test the quality of the software, such as whether the software contains potential security vulnerabilities (Arcuri & Briand, 2014). Many such algorithms, however, are randomized search algorithms (Harman & McMinn, 2010) that do not enumerate all potential execution branches of a software because of the prohibitively expensive computational cost (Arcuri & Briand, 2014).<sup>3</sup> Instead, the algorithms randomly pick the branches to follow at each run. As a result, the algorithm may produce different outputs from



one run to another, even with the exact same input data. The use of such a randomized algorithm can introduce a significant amount of extraneous variance. Fortunately, the nature of the variance is clear—it is affected by nothing but chance, which, importantly, has no interaction with other variables of interest in the study (e.g., personality traits of the engineers who wrote the code or practices of the firm that owns the code). Because the nature of the error is clear to the researcher and is irrelevant to the study variables, its effect on the finding of the study can be made (asymptotically) close to zero by either running the algorithm multiple times to obtain a more reliable output or having a sufficiently large input data set that allows the effect of interest to surface even in the presence of random noise introduced by the randomized algorithm.

*Malignant errors.* The nature of some errors in algorithmic outputs can be unknown to the researchers. We consider these malignant errors. Even though algorithms like deep neural networks (Krizhevsky, Sutskever, & Hinton, 2012) have achieved tremendous success for some information-extraction tasks in practice, many of the state-of-the-art algorithms in this category have been found to produce such highly expressive models that how these models function for a given data set becomes hard to interpret (Bau, Zhou, Khosla, Oliva, & Torralba, 2017). Moreover, how the performance of these models *generalizes* to other input data sets becomes hard to predict even by experts (Szegegy et al., 2013). Such opaqueness in the information-extraction algorithms may pose significant validity threats. For example, a recent study on the performance of several popular commercial algorithms for facial analysis (e.g., those used for gender classification) found that the error rate of the algorithm varies significantly between genders and skin colors—with an error rate of 34.7% for darker-skinned females but only 0.8% for lighter-skinned males (Buolamwini & Gebru, 2018). If a researcher uses such algorithms in a study that involves gender or skin color as a study variable but is oblivious to the uneven errors produced by the algorithms, then the study findings could confound effects that are of theoretical interest (e.g., gender difference in job interview behaviors) with those resulting from the method factor, or they could even purely reflect an artifact of the algorithm performance.

One could arguably attribute malignant errors to a researcher's "blind trust" on the performance of an algorithm or a lack of motivation to study how the algorithm works, just like how some researchers used to blindly run structural equation modeling software without understanding assumptions of the statistical methods and the default setup of the software program (McIntosh, Edwards, & Antonakis, 2014). While the similarity is readily observable, we would like to note that with the fast advance of computer algorithm design, especially in the field of machine learning, even experts in the field may not be able to adequately explain why a state-of-the-art algorithm performs well on some data sets but not as well on others (Bau et al., 2017; C. Zhang, Bengio, Hardt, Recht, & Vinyals, 2017). This special characteristic of algorithmic error makes it even more important for researchers who use the algorithms to be aware of the potential occurrence of malignant errors and to take such errors into account in the research process whenever possible.

*Potential solutions.* To address these potential threats, researchers should extensively test the information-extraction algorithms being used to understand the potential errors produced by the algorithms and whether such errors reduce the statistical power to detect the effects of

theoretical interest or, even worse, distort the data to produce incorrect conclusions of cause and effect. Errors attributed (only) to chance may be mitigated by a larger input data set or, in the case of a randomized algorithm, by running the algorithm multiple times. As for errors inherent in the algorithm design, the burden rests on the researcher who uses the algorithm to carefully examine the distribution of the algorithmic errors and their relationships with the study variables. Remedies here generally require extensive performance testing of the algorithms being used, with the understanding that there is no single silver-bullet performance test that can certify the use of an algorithm will be error free (Ng, 1997; C. Zhang et al., 2017). For example, one may compare multiple competing algorithms to cross-check each other's errors. One could also ask human experts to examine a small sample of algorithmic input/output pairs to identify potential problems (Ribeiro, Singh, & Guestrin, 2016). To avoid the uneven performance of an algorithm across subpopulations, the sample may need to be generated in a stratified manner to cover the full range of the variable that may interact with the algorithm (e.g., equal representation from all races and genders when testing a gender identification algorithm).

### *Threats From a Large Parameter/Procedure Space for an Algorithm*

It is not uncommon to see an information-extraction algorithm applied on organic data starting with preprocessing steps (i.e., “data cleaning”), like “excluding users with fewer than 2 posts in the past year due to inactivity.” Taking these steps implies a choice of *parameter* (e.g., “2” posts instead of “3”) as well as a choice of *procedure* (e.g., filtering nonactive users based on the number of posts rather than the number of friend connections made). Although the number of choices for each parameter/procedure might be limited (e.g., “1 to 3” posts), the overall parameter/procedure space of all possible choices grows exponentially with the number of parameters/procedures and quickly becomes astronomical even for a simple information-extraction algorithm, considering the multiple steps and parameters an algorithm often requires.

*Validity threats.* The existence of parameter selection and design choices made by the researcher—even when made with no sound justification—is not in and of itself a validity threat. After all, many complex algorithms, such as those in machine learning, require carefully tuned parameters and procedures to function properly, yet such tuning is often carried out manually without clear theoretical guidance (Bergstra, Yamins, & Cox 2013; C. Zhang et al., 2017). What truly threatens robustness (or even replicability) of the findings is just like what social science researchers are often warned: If one attempts too many unplanned (post hoc) analyses on the same data, then simply by chance, some tests will result in false positives (Nelson et al., 2018). Following the same logic, a large parameter/procedure space could threaten robustness of the results if a researcher is motivated to test many parameter/procedure choices simply to generate a set of measures that will hit a false positive by chance (Hofman, Sharma, & Watts, 2017). In this case, the line between testing many parameter/procedure choices to ensure good algorithmic performance and cherry-picking the test results can be blurred.

*Potential solutions.* One might find similarities between these validity threats and those caused by *p*-hacking (Nelson et. al., 2018), which is the practice of repeated attempts of modifying statistical models over the same data until a researcher finds statistically significant



results. While both types of threats ultimately result in false positives in research findings, there are subtle yet important differences between the remedies for the two. For example, pre-registration has been suggested as a remedy for *p*-hacking (e.g., Nelson et al., 2018), as it requires a researcher to fix the statistical model for hypothesis testing *before* collecting the data. In contrast, requiring the same pre-registration for algorithmic parameters can be challenging for two main reasons.

One is *when* the pre-registration should take place. Adopting the same rule as traditional experiment design (i.e., pre-register before seeing the data) is challenging because without seeing the data, a researcher often cannot even decide what procedures are needed for the processing of data (note the feedback loop in Figure 1). For example, without knowing (a) whether the data set has a large number of cells with missing values, (b) how many different missing values there are, and (c) the distribution of such missing values, it is difficult for a researcher to decide whether to simply discard all cases with missing values (Daniel, Kenward, Cousens, & De Stavola, 2012), adopt predictive estimation models for the missing values (Cambronero, Feser, Smith, & Madden, 2017), or apply more complex machine-learning algorithms to account for missing values (Zhu, Ghahramani, & Lafferty, 2003) when most values of an attribute are missing. In theory, the researcher could consider all possible cases and pre-register all corresponding strategies. But given the numerous uncertainties about data quality that could require dedicated preprocessing (Rahm & Do, 2000), it is extremely difficult, if not impossible, for a researcher to anticipate all contingency strategies and register each one of them before seeing the data.

Moreover, if the design of the pre-registered process itself cannot be properly evaluated, then threats to the validity of the research could remain even if the pre-registered process is faithfully carried out. While this may not be of significant concern in disciplines like social psychology, where researchers (or readers, reviewers, etc.) can often properly judge the validity of the pre-registered research processes (e.g., measures and statistical analyses used), so far the computer science research community has not yet established common principles, standards, or guidelines on how to judge the appropriateness of the parameter/procedure-tuning process (i.e., how much tuning is deemed too much). There have been numerous attempts of such in different subfields of computer science, such as in data mining (Keogh, Lonardi, & Ratanamahatana, 2004), machine learning (Neyshabur, Bhojanapalli, McAllester, & Srebro, 2017), database systems (Weikum, Moenkeberg, Hasse, & Zaback, 2002), and NLP (Howard & Ruder, 2018). Yet none of these attempts has risen to the level of widely adopted methodological standards (cf. Lipton & Steinhardt, 2018). Without a clear guideline to critique the pre-registered process, it is unclear whether the validity threat could be fully addressed even with pre-registration.

Since the threats from a large parameter/procedure space involve both data and parameter/procedure selection, the remedy is also twofold. First, considering that certain features can be specific to a given data set, a researcher may use part of the input data set as a *tuning sample* to tune the information-extraction algorithm and reserve the rest of the input data set as a *holdout sample* for testing the algorithm after tuning is done. This way, data used for tuning and testing become separated, effectively addressing the false-positive concern (see an example in Campion et al., 2016). It should be noted, however, that the holdout sample is highly similar to the tuning sample as both were obtained at the same time in the same context, which can limit generalizability to other data sets from different settings (Sussmann & Robertson, 1986). More importantly, while a holdout sample has been shown as an effective

tool for preventing the tuned parameters from “overfitting” (Ng, 1997) the data (see illustration in Putka, Beatty, & Reeder, 2018), researchers should still exercise caution when using a holdout sample because it is the proper usage of a holdout sample in parameter tuning, not the mere construction of a holdout sample itself, that addresses the validity threat stemming from a large parameter space.

For example, a holdout sample can be used only once, which is often neglected in practice. Consider a researcher who first validates the result over a holdout sample, then uses knowledge of the validation result to form another hypothesis and tests this second hypothesis in the same holdout sample. This practice is not uncommon in conceptual replication studies where researchers use the same data set to validate an existing finding first and then explore (in a post hoc fashion) related relationships (e.g., a mediation or moderating relationship). Yet doing so would make the original holdout sample no longer qualified as a validation tool for the second hypothesis because the sample supports the hypothesis *by design* (Dwork, Feldman, Hardt, Pitassi, Reingold, & Roth, 2015b). As another example, recent research in machine learning, a field where holdout samples are universally used for performance testing, has found widespread abuse of holdout sample usage, which has created significant overfitting problems and artificially low error rates that cannot be replicated on other data sets (Cawley & Talbot, 2010; Reunanen, 2003). Fortunately, when collecting new data from multiple samples is costly or impossible, recent developments in theoretical computer science suggest that using a randomized approach can enable the same holdout sample to be reused (see further explanations in Dwork et al., 2015b; Dwork, Feldman, Hardt, Pitassi, Reingold, & Roth, 2015a).

The other potential remedy is to investigate the parameter/procedure space to detect signs of questionable practices. There are two types of solutions here. One is to institute editorial policies that require researchers to fully disclose the parameters or procedures attempted before reaching the final design. While there can be some deterrence effect of such policies, as discussed earlier, it is important to note that the larger machine-learning research community has not yet established common guidelines on how to judge the appropriateness of the parameter/procedure-tuning process. The other type of solution is running robustness checks. For example, one can randomly sample other parameter/procedure combinations that may make changes to the input data set and test whether the research finding remains consistent. The premise here is that if the finding changes after adjusting the parameter for data filtering (e.g., “excluding users with fewer than 1 [rather than 2] posts”) or the procedure (e.g., filtering nonactive users based on the number of friend connections made rather than the number of posts), then the finding is more likely to be an artifact of the specific parameter/procedure selection than a robust relationship of theoretical relevance. While this method is feasible in simple procedures, like the aforementioned example, it is still an open question how it can be applied to more complex algorithms for which the relationship between parameter settings and algorithm performance remains opaque even for experts (Bergstra et al., 2013; Szegedy et al., 2013). There is a chance that only one set of parameters/procedures works well for a specific data source, and it could be premature to attribute the failure of other parameter settings to the lack of validity of a research finding, considering the opaqueness of the complex algorithms. Another issue to be considered is that multiple algorithm specifications with variations in parameter values may generate highly similar outputs and thus yield similar findings, which resembles the issue of model selection uncertainty in statistical analyses

(Burnham & Anderson, 2002). Researchers should be cautious about interpreting such equifinality in algorithm specification as evidence of lack of robustness of the algorithm used.

## Threats to Validity From the Opaque Data Generation Process

A key difference between the usage of “organic” and “designed” data is readily spelled out in the names—researchers have much less control on how organic data were generated. This opaqueness of the data generation process poses special challenges for researchers to properly make inferences about the relationships between operationalizations and constructs (i.e., construct validity) and about the relationships among variables (i.e., internal validity). While our discussions in this section hence focus on these two validity components, it is important to note that the opaqueness of the organic data generation process can also incur threats to external validity and statistical conclusion validity, because (a) external validity and statistical conclusion validity are closely related to construct validity and internal validity (Shadish et al., 2002); (b) the usage of organic data often demands the usage of information-extraction algorithms, the issues of which we elaborated earlier; and (c) factors in the data generation process that differ from one population to another and yet are unknown to the researchers because of the opaqueness of the processes can threaten external validity as well as internal validity and construct validity. This last issue needs to be carefully addressed in observational studies in order to establish causality. The existence of unknown factors that drive observed data threatens internal validity because it is difficult to ensure that the observed effect cannot be attributed to any confounding factors (which fails to satisfy the critical condition for establishing causality; Pearl, 2009). When these unknown data generation factors also vary from context to context, the external validity would be affected, as well.

### *Threats to Construct Validity*

Two steps in the processing of organic data can incur threats to construct validity: (a) ad hoc data prefiltering and (b) post hoc measurement design.

*Ad hoc data prefiltering.* The large volume of organic data, along with the fact that they are not generated for any specific research goal and therefore contain a large amount of irrelevant data, makes it necessary to filter the data before using them for research purposes. For example, to use data from Twitter, instead of including every tweet in the statistical analyses, a practical solution is to first filter tweets based on preset criteria, such as keywords, and then analyze only those tweets that pass the filter. Fully validating a filter would require thoroughly examining the data set, which would defeat the purpose of setting up filters. Therefore, many existing studies have resorted to ad hoc filter designs, such as an arbitrary selection of keywords subjectively decided by the researchers to be “relevant” to the topic of interest. Unfortunately, such ad hoc designs likely introduce significant biases to the subsequent measurements of constructs using the filtered data. Past studies found that, for the same topic, the sets of keywords used by different individuals may have little overlap with each other but correlated with individuals’ opinions or past experiences about the topic (H. Zhang, Hill, & Rothschild, 2016). For example, whether one uses the hashtag #oscar or #neveroscar when discussing the Academy Awards ceremony depends on the person’s

opinion on the event. If a filter includes only #oscar but not #neveroscar, the selected tweets could be systematically biased. In addition, it can be difficult to ensure a proper filter design without knowing what data might appear in the input data set. For example, an intuitive filter to use for identifying tweets relevant to different cities is city names. However, such a filter may produce a data set with much more noise from tweets including city names like Houston (e.g., including tweets mentioning “Whitney Houston” in the filtered data set) versus others (e.g., Indianapolis). In all these cases, the filter is designed according to the researcher’s *assumptions* about the data generation process that unfortunately miscategorized important data. When the filtered data set systematically contains irrelevant data or misses parts of the conceptual space, the measurements developed later can be contaminated or deficient.

*Measurement design.* The proper operationalization of a construct is often more challenging over organic data not generated by research design than traditional observational studies using established scales. For example, when researchers need to capture the similarity between two Twitter accounts as perceived by other Twitter users, it is impractical for researchers to directly elicit a similarity score from Twitter users using a survey. Instead, they have to resort to inference-based operationalization, such as measuring the number and percentage of users who follow both accounts (e.g., Culotta & Cutler, 2016). However, such a seemingly direct correspondence between the construct and the operationalization can often include contaminated parts in the measure. For example, a study of social link creation time on Twitter found that nearly half of following events happen within a day after the follower joined Twitter (Meeder, Karrer, Sayedi, Ravi, Borgs, & Chayes, 2011). This suggests that two accounts may acquire a large number of common followers simply by being in the news and hence appearing in Twitter’s “Trends for you” or “Who to follow” lists on the same day (Gupta, Goel, Lin, Sharma, Wang, & Zadeh, 2013). In fact, research has shown that the rationale underlying a user’s “following” action differs significantly from one account being followed to another.<sup>4</sup> Therefore, the number of common followers of two accounts may not be a proper measure of the similarity between these two accounts.

*Potential solutions.* A general recommendation from Landers et al. (2016) is to develop a data source theory to guide data processing and measurement development in organic data studies. As demonstrated by Landers et al., the data source theory must be clearly communicated to readers, and when hypotheses derived from the data source theory are not supported, data-processing procedures need to be revised and implications on findings need to be fully discussed. More specifically, in the case of keyword-based data filtering, instead of only relying on the preselected, subjectively decided set of keywords, the researcher can test the comprehensiveness of the filter by using a technique called “bootstrapping” in computer science literature (e.g., M. Zhang, Zhang, & Das, 2013). Commonly used in information retrieval, the bootstrapping technique first identifies additional keywords from those data retrieved by the preselected keywords and then checks whether a significant amount of additional data can be retrieved from discarded data based on the new keywords. This process can be carried out iteratively until no additional data can be revived from the discarded pile. In the case of measurement design, given the researcher’s lack of control over the data generation process and the difficulty of accessing data sources (e.g., individual human users of Twitter), many conventional construct validity tests, such as the multitrait-multimethod model, may not be

directly applicable to organic data. Nonetheless, this does not mean validating a measure is infeasible for organic data (see examples of validation in Sajjadi et al., 2019; Wang et al., 2016). Quite the contrary, conceptual replications using slightly varied operationalizations for the same constructs can still alleviate many threats to construct validity. For example, in addition to counting the total number of common followers, one could measure the common followers in each country separately, the common followers who interact with both friends through retweets, or the common words in tweets mentioning the two brands. If there is an agreement among multiple similar measures, the converging evidence could alleviate concerns about the threats to construct validity stemming from the mono-operationalization approach (Shadish et al., 2002).

### *Threats to Internal Validity*

Although the principles of experimental design remain the same for studies using organic data as in traditional experiments, in practice, many experiments using organic data are conducted on complex technological platforms with opaque designs (e.g., online social networks, movement-tracking sensors). This opaqueness often makes it impossible or prohibitively expensive to enable proper experimental manipulations. For example, researchers often do not have direct controls on what recommendations or advertisements a user sees on a website. Instead, they often resort to alternatives such as natural experiments (Sismeiro & Mahmood, 2018), which leverage natural incidents, like website outages, as the manipulation. Or they may leverage techniques like the causal inference approach developed by Pearl (2009), which attempts to infer causal relationships from observational data based on tools like causal diagrams. It is well known that the lack of control on manipulation and randomization of participants makes it difficult to exclude alternative explanations for the observed covariations, introducing concerns on internal validity. Making it even more challenging to draw causal inferences, the black-box design of technological platforms can introduce hidden linkages between different variables captured in the same study, thereby confusing cause and effect with the internal works of the technological platforms. Such threats may occur in the following three forms.

*Links between features.* Separate, seemingly independent features of a technological platform may be linked at the back end. If a researcher is unaware of such linkages, then he or she might incorrectly attribute the observed covariation between two variables to a treatment effect when in fact that covariation is caused by the linkage inside the platform design. For example, researchers may be interested in testing how the new connections a user forms on a website (e.g., LinkedIn) affect the user's behavior (e.g., searching or applying for job vacancies). While the new-connection feature may appear separate from the job post-following feature of the website, many such websites provide recommendations for both interuser connections and posts about products/corporations based on shared factors, such as a user's past browsing history. In this case, the covariation one observes between the new connection and the change of job search behaviors may be both caused by the hidden factor of platform-generated recommendations. This is a typical example of how a hidden link between features of a technological platform threatens the internal validity of causal inferences from an experiment.

*Dynamic changes.* Features of online platforms (e.g., online social networks and e-commerce websites) evolve over time, often at a fast speed. However, the exact nature of these changes may not be announced timely or at all. The Android app of Instagram, for example, was updated 286 times from its first release in April 2012 to May 2018, averaging 3.86 times per month (Uptodown.com, 2018). While some of these updates were minor security fixes, others introduced major revisions to how the platform functioned (Segarra, 2018). Such major revisions could change what feeds into the measurements used in a study and thereby lead to incorrect inferences about the relationship of theoretical interest. For example, Apple has revised several times how the average review score is computed for an app in its iOS App Store (Dillet, 2017). The score started off as the average of all reviews for the app, was subsequently changed to be the average of reviews for the current version only, and then, in 2017, became much more fluid as Apple began to allow a developer to choose whether to reset the average review score when an app is updated. This last change could artificially increase the average review score, as a developer can choose whether to reset the score based on the trend of the last batch of reviews. If the current review score were used as an indicator of product performance or performance of the research-and-development team in a study, changes in the score over time could be mistaken as only reflecting a certain type of treatment effect when, in fact, it also reflected changes in the measurements (which may have captured different theoretical constructs).

*Link between platforms.* With the increasing prevalence of cloud computing, many seemingly unrelated technological platforms operated by different organizations may now have hidden links attributable to their common cloud service providers. For example, many websites today use one or more content delivery networks (CDNs), a collection of geographically distributed data centers and servers, to quickly deliver content to their users around the globe. When one of the major CDNs (e.g., Akamai, Amazon CloudFront, Azure CDN) suffers a service outage, websites that use the same service will be simultaneously affected. Moreover, since many websites use more than one CDN, the affected users may be limited to isolated geographic regions or Internet service providers, making such links difficult to detect. Not being aware of such a link would be problematic when website outages are used as the basis of “natural experiments” in research aiming to understand how users move from one website to another (Sismeiro & Mahmood, 2018). If a researcher is unaware of the common CDNs used by multiple websites, then he or she might incorrectly attribute the observed covariation of traffic drops to the treatment effect (e.g., features manipulated on one website of theoretical interest leading users to another) when in fact the covariation is caused by the technical issue of the underlying shared CDN.

*Potential solutions.* An ideal solution to the above-described problem is, of course, to thoroughly “unbox” the design of the technological platform. However, this is often infeasible due to the sensitivity of such designs. For instance, the recommendation algorithm is often considered a trade secret by websites such as Amazon. As such, the researcher has to be diligent to uncover the potential hidden links and dynamic changes, and properly institute checks to guard against the threats to internal validity. For example, to guard against potential validity threat from between-feature links in the earlier example, it is important to carefully document all data a user may be exposed to on the website and check through empirical



tests to identify which of them may be affected by the treatment condition. Similarly, one can use tools like CDN finders (e.g., [whatsmycdn.com](http://whatsmycdn.com)) to detect CDNs used by a website. The continuous monitoring of measured variables to detect sudden changes, using multiple operationalizations when possible, along with cross-checking changes with known system updates, can alleviate threats from dynamic changes. Nevertheless, none of these solutions can completely eliminate threats from the black-box design to internal validity because, as the name “black box” suggests, the platform provider might make unannounced changes that are extremely difficult, if not impossible, to detect (as an example, see technical work on how websites can institute hard-to-detect changes to thwart external attempts on reverse engineering its design; M. Zhang, Zhang, & Das, 2012). Despite the challenge, the diligent checks by researchers can alleviate many common concerns and boost confidence in the robustness of the findings.

## Final Remarks

In this commentary, we discussed two broad types of validity threats concerning research using organic data, one stemming from the opaqueness of the organic data generation process and the other from the need of using automated information-extraction algorithms to process such data. We also shared some potential remedies. While much of our discussion focused on the potential pitfalls of using organic data, we would like to conclude the commentary by cautioning against a “defeatist attitude,” that is, one that completely dismisses the scientific value of organic data for research solely due to the validity concerns. This attitude could be detrimental to advancing research in any field, especially management, given the importance of organic data to today’s businesses and the potential for organic data to complement data from design-driven studies. We believe, instead of abandoning organic data (which is to desert a gold mine of valuable information and knowledge), the research community should invest more in the proper understanding on the extent of, causes of, and remedies for validity threats in the workflow of any research using organic data. It is our hope that this commentary demonstrates the pressing need of such research and inspires more future studies in this direction.

## ORCID iD

Heng Xu  <https://orcid.org/0000-0001-5642-6543>

## Notes

1. Note that the robustness issues discussed in this commentary overlap with, but are not entirely identical to, issues covered by these related terms, such as reproducibility and replicability. For example, recent studies (Chapman, Benedict, & Schiöth, 2018) found that certain characteristics (e.g., gender) of the experimenter who interacts with subjects in a clinical trial could have significant effects on the research findings. While such experimenter-related issues are generally covered under the umbrella of reproducibility or replicability, they are not as relevant in research using organic data and hence are not the focus of this commentary.

2. Both U.S. Census data and e-mail traces are generally considered *archival data* in management research. See a review of different types of archival data used in management research in Barnes, Dang, Leavitt, Guarana, and Uhlmann (2015).

3. For example, an *if* statement, such as “*if* (*age* < 30) [code-block-1] *else* [code-block-2],” creates two branches. Which branch will be activated depends on input values (*age*) that may not be known prior to execution. This type of nested *if* statement creates an exponential number of possible execution paths, leading to a prohibitively high computational cost for enumerating all possibilities.

4. For example, see the sharp contrast between the engagement rates of the followers of @chefsymon and those of @zagat at <https://moz.com/blog/social-authority>.

## References

- Aguinis, H., Ramani, R. S., & Alabduljader, N. 2018. What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, 12: 83-110.
- Aho, A. 1990. Algorithms for finding patterns in strings. In J. van Leeuwen (Ed.), *Handbook of theoretical computer science: Volume A. Algorithms and complexity*: 255-300. Cambridge, MA: MIT Press.
- Aral, S., & Walker, D. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 57: 1623-1639.
- Arcuri, A., & Briand, L. 2014. A hitchhiker’s guide to statistical tests for assessing randomized algorithms in software engineering. *Software Testing, Verification and Reliability*, 24: 219-250.
- Barnes, C. M., Dang, C. T., Leavitt, K., Guarana, C. L., & Uhlmann, E. L. 2015. Archival data in micro-organizational research: A toolkit for moving to a broader set of topics. *Journal of Management*, 44: 1453-1478.
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*: 3319-3327. Piscataway, NJ: IEEE.
- Bergstra, J., Yamins, D., & Cox, D. D. 2013. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the International Conference on Machine Learning*, 28: 115-123.
- Blodgett, S. L., & O’Connor, B. 2017. *Racial disparity in natural language processing: A case study of social media African-American English*. Available from arXiv. (arXiv:1707.00061)
- Braun, M. T., & Kuljanin, G. 2015. Big data and the challenge of construct validity. *Industrial and Organizational Psychology*, 8: 521-527.
- Buolamwini, J., & Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Conference on Fairness, Accountability and Transparency*, 81: 77-91.
- Burnham, K. P., & Anderson, D. R. 2002. *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer Science + Business Media.
- Cafarella, M. J., Madhavan, J., & Halevy, A. 2009. Web-scale extraction of structured data. *ACM SIGMOD Record*, 37: 55-61.
- Cambroner, J., Feser, J. K., Smith, M. J., & Madden, S. 2017. Query optimization for dynamic imputation. *Proceedings of the VLDB Endowment*, 10: 1310-1321.
- Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. 2016. Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, 101: 958-975.
- Cawley, G. C., & Talbot, N. L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11: 2079-2107.
- Chapman, C. D., Benedict, C., & Schiöth, H. B. 2018. Experimenter gender and replicability in science. *Science Advances*, 4: e1701427.
- Culotta, A., & Cutler, J. 2016. Mining brand perceptions from twitter social networks. *Marketing Science*, 35: 343-362.
- Daniel, R. M., Kenward, M. G., Cousens, S. N., & De Stavola, B. L. 2012. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21: 243-256.
- Davidov, D., Tsur, O., & Rappoport, A. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Conference on Computational Natural Language Learning*: 107-116. Stroudsburg, PA: Association for Computational Linguistics.
- Dillet, R. 2017. The new iOS App Store lets devs choose whether or not to reset ratings when updating. *TechCrunch*. Retrieved from <https://techcrunch.com/2017/06/07/ios-app-developers>
- Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. L. 2015a. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Annual Symposium on Theory of Computing*: 117-126. New York: ACM.

- Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. 2015b. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349: 636-638.
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. 2016. What does research reproducibility mean? *Science Translational Medicine*, 8: 341ps12.
- Groves, M. R. 2011. Three eras of survey research. *Public Opinions Quarterly*, 75: 861-871.
- Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., & Zadeh, R. 2013. WTF: The who to follow service at Twitter. In *Proceedings of the International Conference on World Wide Web*: 505-514. New York: ACM.
- Harman, M., & McMinn, P. 2010. A theoretical and empirical study of search-based testing: Local, global, and hybrid search. *IEEE Transactions on Software Engineering*, 36: 226-247.
- Hofman, J. M., Sharma, A., & Watts, D. J. 2017. Prediction and explanation in social systems. *Science*, 355: 486-488.
- Howard, J., & Ruder, S. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*: 328-339. Stroudsburg, PA: Association for Computational Linguistics.
- Keogh, E., Lonardi, S., & Ratanamahatana, C. A. 2004. Towards parameter-free data mining. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 206-215. New York: ACM.
- King, G., & Lowe, W. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57: 617-642.
- King, G., & Zeng, L. 2007. When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly*, 51: 183-210.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*: 1097-1105. Red Hook, NY: Curran Associates.
- 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 psychology research. *Psychological Methods*, 21: 475-492.
- Lipton, Z. C., & Steinhardt, J. 2018. *Troubling trends in machine learning scholarship*. Available from arXiv. (arXiv:1807.03341)
- McFarland, L. A., & Ployhart, R. E. 2015. Social media: A contextual framework to guide research and practice. *Journal of Applied Psychology*, 100: 1653-1677.
- McIntosh, C. N., Edwards, J. R., & Antonakis, J. 2014. Reflections on partial least squares path modeling. *Organizational Research Methods*, 17: 210-251.
- Meeder, B., Karrer, B., Sayedi, A., Ravi, R., Borgs, C., & Chayes, J. 2011. We know who you followed last summer: Inferring social link creation times in Twitter. In *Proceedings of the International Conference on World Wide Web*: 517-526. New York: ACM.
- Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. 2011. Understanding the demographics of Twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*: 554-557. Menlo Park, CA: AAAI Press.
- Nelson, L. D., Simmons, J., & Simonsohn, U. 2018. Psychology's renaissance. *Annual Review of Psychology*, 69: 511-534.
- Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. 2017. Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, 28: 5947-5956.
- Ng, A. Y. 1997. Preventing "overfitting" of cross-validation data. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*: 245-253. Burlington, MA: Morgan Kaufmann.
- Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349: aac4716.
- Ott, M., Cardie, C., & Hancock, J. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st International Conference on World Wide Web*: 201-210. New York: ACM.
- Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3: 96-146.
- Putka, D. J., Beatty, A. S., & Reeder, M. C. 2018. Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, 21: 689-732.
- Rahm, E., & Do, H. H. 2000. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23: 3-13.
- Reunanen, J. 2003. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3: 1371-1382.

- Ribeiro, M. T., Singh, S., & Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 1135-1144. New York: ACM.
- Sajjadiani, S., Sojourner, A., Kammeyer-Mueller, J. D., & Mykerezi, E. 2019. Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, 104: 1207-1225.
- Schwaiger, J. M., Lang, M., Ritter, C., & Johannsen, F. 2016. Assessing the accuracy of sentiment analysis of social media posts at small and medium-sized enterprises in southern Germany. *Proceedings of the European Conference on Information Systems*, 107: 1-17.
- Segarra, L. M. 2018. Instagram is making a big change and now new posts will show up first again. *Time*. Retrieved from <http://time.com/5210976/instagram>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Sismeiro, C., & Mahmood, A. 2018. Competitive vs. complementary effects in online social networks and news consumption: A natural experiment. *Management Science* 64. <https://doi.org/10.1287/mnsc.2017.2896>
- Sussmann, M., & Robertson, D. U. 1986. The validity of validity: An analysis of validation study designs. *Journal of Applied Psychology*, 71: 461-468.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. 2013. *Intriguing properties of neural networks*. Available from arXiv. (arXiv:1312.6199)
- Uptodown.com. 2018. *Instagram old versions*. Retrieved from <https://instagram.en.uptodown.com/android/old>
- Wang, W., Hernandez, I., Newman, D. A., He, J., & Bian, J. 2016. Twitter analysis: Studying US weekly trends in work stress and emotion. *Applied Psychology: An International Review*, 65: 355-378.
- Weikum, G., Moenkeberg, A., Hasse, C., & Zabback, P. 2002. Self-tuning database technology and information services: From wishful thinking to viable engineering. In *Proceedings of the 28th International Conference on Very Large Databases*: 20-31. Hong Kong: VLDB Endowment.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. 2017. *Understanding deep learning requires rethinking generalization*. Paper presented at the International Conference on Learning Representations, Toulon, France.
- Zhang, H., Hill, S., & Rothschild, D. 2016. Geolocated Twitter panels to study the impact of events. In *2016 AAAI Spring Symposium Series*: 318. Menlo Park, CA: AAAI Press.
- Zhang, M., Zhang, N., & Das, G. 2012. Aggregate suppression for enterprise search engines. In *Proceedings of the ACM SIGMOD Conference on Management of Data*: 469-480. New York: ACM.
- Zhang, M., Zhang, N., & Das, G. 2013. Mining a search engine's corpus without a query pool. In *Proceedings of the ACM Conference on Information and Knowledge Management*: 29-38. New York: ACM.
- Zhu, X., Ghahramani, Z., & Lafferty, J. D. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*: 912-919. Menlo Park, CA: AAAI Press.