American Psychological Association

American Psychologist


0003-066X/18/$12.00

2018, Vol. 73, No. 7, 899–917
http://dx.doi.org/10.1037/amp0000190

# Big Data in Psychology: A Framework for Research Advancement

Idris Adjerid and Ken Kelley
University of Notre Dame

The potential for big data to provide value for psychology is significant. However, the pursuit of big data remains an uncertain and risky undertaking for the average psychological researcher. In this article, we address some of this uncertainty by discussing the potential impact of big data on the type of data available for psychological research, addressing the benefits and most significant challenges that emerge from these data, and organizing a variety of research opportunities for psychology. Our article yields two central insights. First, we highlight that big data research efforts are more readily accessible than many researchers realize, particularly with the emergence of open-source research tools, digital platforms, and instrumentation. Second, we argue that opportunities for big data research are diverse and differ both in their fit for varying research goals, as well as in the challenges they bring about. Ultimately, our outlook for researchers in psychology using and benefiting from big data is cautiously optimistic. Although not all big data efforts are suited for all researchers or all areas within psychology, big data research prospects are diverse, expanding, and promising for psychology and related disciplines.

*Keywords:* big data, data science, machine learning, instrumentation



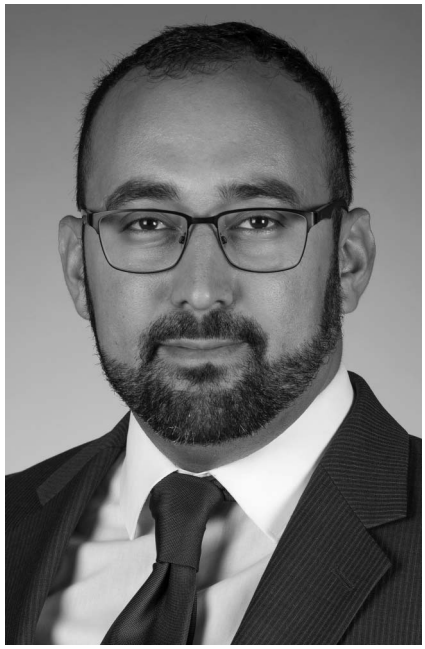
Large and dynamic data sets now exist or can be collected that capture granular and diverse characteristics about thousands, and in some cases millions, of individuals at a single point in time and longitudinally. These data are obtained primarily with the use of large digital platforms, have the potential to inform important questions, and have fueled work using "big data" in commercial settings (Chen, Chiang, & Storey, 2012), as well as some academic areas, particularly computer science (e.g., Chen et al., 2004; Somanchi, Adhikari, Lin, Eneva, & Ghani, 2015). Over the last several years, scholars in psychology have become more interested and engaged in exploring the potential of big data from digital platforms to inform important questions in the field (Jaffe, 2014). A few recent and high-profile research efforts illustrate advances in psychological insight made possible by the big data generated by popular digital platforms. Youyou, Kosinski, and Stillwell (2015), for example, used Facebook "likes" data from 90,000 study participants to create predictive models for inferring individual personality characteristics. Muchnik, Aral, and Taylor (2013) used a randomized field experiment on a popular news website in an attempt to understand the impact of social influence on ratings of news stories. Undoubtedly, these large data collection efforts generated by interactive digital platforms provide a way of understanding psychological constructs and processes that has been impractical, if not impossible, until only very recently (Jaffe, 2014).

This promise for psychology to make progress by leveraging big data, while exciting, also raises significant questions and concerns for researchers, particularly those with little or no experience with collecting, preparing, and analyzing big data—and recent work suggests that such is the case for the majority of researchers in psychology at present (Metzler, Kim, Allum, & Denman, 2016). In our experience, researchers in psychology are often uncertain about exactly how the "big data era" is changing the structure of data available for research and the implication of these changes for their specific questions of interest and methods of choice (e.g., the extent to which big data will shift the research focus in psychology to predictive or exploratory efforts). Moreover, researchers have significant questions about whether big data research is within their reach because of a widening "digital divide" where some select researchers at elite institutions are "Big Data rich" but the majority of researchers are "Big Data poor" (Boyd & Crawford, 2012, p. 674). Moreover, the technical expertise necessary for collecting and organizing big data for use in studies is currently more in line with computer science

This article was published Online First February 22, 2018.

Idris Adjerid and Ken Kelley, Department of Information Technology, Analytics, and Operations, Mendoza College of Business, University of Notre Dame.

Correspondence concerning this article should be addressed to Idris Adjerid, who is now at the Department of Business Information Technology, Pamplin Hall, RM 2058, Virginia Tech, 880 West Campus Drive, Blacksburg, VA 24061. E-mail: iadjerid@vt.edu



**Idris Adjerid**
Photo by
Barbara Johnston

training than with traditional psychological science training. If unaddressed, these issues may manifest as significant barriers to the pursuit of big data research in the broader community of researchers in psychology. In this article, we attempt to break down some of these barriers by raising and addressing a variety of questions and concerns. In so doing, we hope to provide a footing for the "average" researcher in psychology who wishes to engage in big data research, which we believe is a promising but complex landscape.

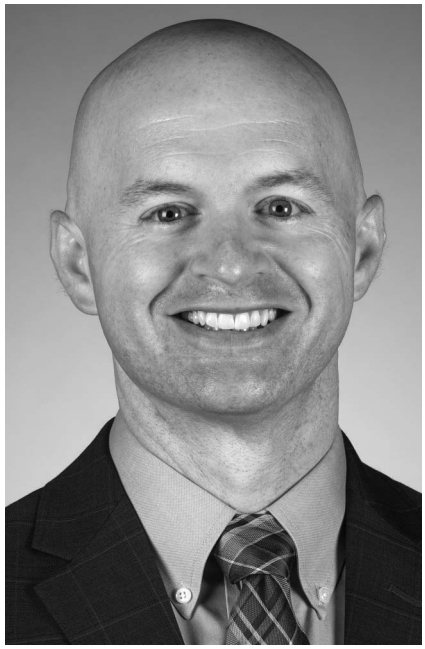
We start by simplifying how the "big data era" is changing the structure of data available for research and argue that highly instrumented digital platforms will have dramatic impacts on the scale of "persons" available for study (sample size, *n*), the novelty and diversity of the variables (variables, *v*) available about these persons, and the ability to observe changes in these variables over many more occasions (time, *t*). At the same time, digital platforms generally collect data indiscriminately, often without any research questions in mind. Thus, the data obtained are often highly unstructured and diverse, and can hold uncertain value for exploring research questions in psychology. With this reality in mind, we discuss the benefits that these changes in available data introduce for psychological researchers, as well as the corresponding complexities and challenges, and point to some pathways for researchers to overcome these challenges. We follow this discussion with a breakdown of the nuanced and diverse research opportunities made possible for psychology by some combination of large sample size (big *n*), a rich set of variables about individuals and/or groups (big *v*), and granular and sustained data collection over time (big *t*). We supplement this breakdown with numerous examples of contemporary research efforts across diverse fields to make more tangible the potential big data efforts that researchers in psychology can pursue. We conclude with a discussion of the ethical and privacy considerations associated with big data research in psychology and provide some final thoughts on the direction of such research in this field.

We provide a number important insights that we hope provide clarity on some of the most pressing questions for researchers contemplating big data research in psychology. First, we highlight that big data research efforts are much more within reach than many researchers realize. Specifically, we argue that big data research goes well beyond the number of participants (i.e., sample size), which has at times been considered to be the primary factor when considering what makes data "big." In addition, we point researchers to works that are starting to narrow the gap between the traditional methodological competencies of psychology and what is needed to navigate the big data landscape. Finally, we highlight that there are a variety of pathways for gaining access to big data for research purposes (scraping data, third party vendors, or crowd-sourcing platforms). It is important that many of these pathways do not require a collaborative commitment from the platform owners, which can be difficult to obtain. Big data are even more accessible if researchers realize that they can craft their own research settings that are instrumented to capture rich and granular data about individuals.

Second, we highlight that opportunities for big data research can take diverse forms and that these diverse forms differ in their fit for various research efforts and goals. For instance, researchers with targeted questions about relationships that may be highly heterogeneous in the population of interest may benefit from observing the real-world behavior of large, diverse samples but may only require a few variables about these individuals. On the other hand, researchers interested in psychometrics and measurement may want to explore how constructs of longstanding interest to psychology (dimensions of personality, need for cognition, motivation, etc.) reveal themselves in disparate data left by users of these platforms (sometimes termed "data breadcrumbs"). The broader point is that big data efforts are diverse and we believe most researchers in psychology can benefit from the opportunities big data present. At the same time, not all big data efforts fit all research contexts or individual researchers, and big data cannot substitute for careful research design and the appropriate consideration of research questions.

## Big Data and Its Impact on "Research As We Know It"

To put big data research in perspective, it is useful to briefly discuss the current state of affairs in psychological research. In traditional psychological research, there contin-



**Ken Kelley**
Photo by
Barbara Johnston

ues to be a focus on a single outcome variable with relatively few explanatory variables. If these variables are measured over time, they are usually measured at highly structured and discrete occasions often specified a priori (e.g., one trip to the laboratory each week for 5 weeks). The vast amount of research design literature in psychology and related disciplines is based on this scenario of research with only a few variables (*v*), captured cross-sectionally or on highly structured occasions (*t*), for (relatively) few research participants (*n*). In fact, a large part of the research design literature attempts to find as *small* a sample size (*n*) as is reasonable to address the specific question of interest (e.g., using a power analysis, which seeks to find the minimum sample size necessary in order to have at least 80% power to detect a truly medium or larger effect). In many ways, this combination of few variables and small sample size has been typical of empirical research in psychology for the last century. The questions answered by such traditional research efforts are purposefully and necessarily limited in scope, often focusing on partitioning variance and estimating effects between specific variables. Unsurprisingly, many research methods employed in traditional psychological research are well known and highly vetted (e.g., *t* tests, analysis of variance [ANOVA], multiple regression, chi-square goodness-of-fit, psychometrics).

Over time, some important and noteworthy limitations of this type of research have emerged that are relevant to big data research. For example, more than 50 years ago it started to become evident that researchers were often using too small a sample size for effective research (e.g., Cohen, 1962). Moreover, traditional research often uses "convenience" samples to test and attempt to validate psychological theories, even though these samples are not representative of the population to which researchers often hope to generalize their findings (Henrich, Heine, & Norenzayan, 2010). In addition, the measurement of highly dynamic constructs (e.g., mood, emotion) in psychology is often too coarse and useful variation in these constructs is often not accurately measured, which has led to the development of intensive longitudinal methods (Csikszentmihalyi & Larson, 2014). Finally, there have been revelations of likely rare but significant research fraud (e.g., Simonsohn, 2013), as well as potentially more widespread practices of data manipulation such as "*p*-hacking" (i.e., reporting only conditions that "worked," and post hoc theorizing; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). Added to the use of sample sizes that are often still too small for robust and replicable findings, these issues contribute significantly to the so called "replication crisis" in psychology (Maxwell, Lau, & Howard, 2015).

We contend that big data emerging from large digital sources will both complement and extend traditional psychological research in the coming decade and beyond. In particular, we believe that change will occur with regard to the methods employed in research, the nature of data limitations, and ethical considerations (e.g., privacy). Before diving into these considerations, however, we first simplify how the "big data era" is changing the structure of data available for research through the lens of two foundational works in research methods. The first is Cattell's "data box" (Cattell, 1946, p. 93; see also Cattell, 1966), in which he classifies methods based on the structure of data and, in its simplest form, organizes data along three dimensions: persons, variables, and occasions. In the context of Cattell's data box, an instrumented world and the big data that it generates will have dramatic impacts on each of these dimensions with increases in the scale of "persons" available for study (sample size, *n*), the novelty and diversity of the variables (variables, *v*) available about these persons, and the ability to observe changes in these variables over many more occasions (time, *t*). Our conception of big data using Cattell's data box, while not identical, parallels other contemporary views of big data which posit that big data can be characterized by three *V*s: volume, variety, and velocity of data (Borgman, 2015).

The second work is Coombs's (1964) *Theory of Data*, in which he notes that formal statistical methods, in search of insight, leverage observations that are selected from a universe of potential observations, and parsed into usable information for use in statistical models (chapter 1). In Coombs's framework, these (raw) observations provide choices for which data to parse into meaningful variables and how to use such variables in research, all of which becomes more complex with big data. Furthermore, data from these digital platforms are often rich but collected indiscriminately, often without any research questions in

mind. Such an issue can result in not only highly unstructured and diverse data, but also data with uncertain value for exploring research questions in psychology. Combining the perspectives of Cattell and Coombs, we argue that big data research will involve many more potential participants and much more information about them. At the same time, these data are less structured and less readily integrated into existing research efforts. In what follows, we consider some of the significant impacts on research in psychology as data becomes "big" along *n*, *v*, or *t*.

## Big "*n*"

The digital platforms that underlie some big data collection efforts can provide access to tens of thousands and in some cases millions of individuals for research. Unlike more traditional data sources and data collection methods (e.g., student populations, face-to-face interviews, laboratory studies, etc.), large-scale digital platforms are highly scalable in their ability to collect data on real-world behaviors for a large number of individuals. In addition, some such platforms offer a way to not only observe these individuals but to introduce interventions or communicate with them cost-efficiently and at scale unprecedented for individual researchers until very recently. Coupling these capabilities with the fact that many of these platforms have enjoyed high rates of adoption and use, results in digital platforms' potentially providing access to large swaths of the (online) population, capturing real-world outcomes and behaviors of interest to psychology, and providing mechanisms for interacting with these individuals as well as altering their decision environment. These larger samples sizes, across a more diverse set of individuals, allow for the detection of specific effects with a high degree of precision, and, more generally, allow for the estimation of complex statistical models. Moreover, access to broad swaths of the online population also has the potential to facilitate research samples that are considerably more representative of their target population (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Shannon, Andrew, & Duggan, 2016; Ramo & Prochaska, 2012), though even some groups will remain elusive even in a more diverse set of individuals overall. Of course, researchers still need to consider the sample selection concerns around which type of individuals respond to recruitment efforts on these digital platforms.

## Big "*v*"

The large digital platforms that underlie the emergence of big data will also have a considerable impact on the variety of variables available for research. Whereas large sample sizes are driven by the vast uptake and participation on digital platforms and instrumentation, the increase in variety of variables available for research is driven by the rich nature of the interactions on these platforms and the ability of these platforms to measure this behavior at a granular level. Individuals online can upload images, write, edit, and delete posts on social networks, up-vote/down-vote stories, share and consume various content (articles, videos, movies, etc.), search for certain things, and peruse various products then decide to purchase (or not). All of this behavior is observed at some level on these platforms, making for a diverse set of variables that can be derived about platform users. The ability to capture all of these interactions makes even nonevents equally interesting (e.g., what a user *did not* click). Ultimately, big "*v*" results in many more potential measures of individual behavior available for consideration by researchers and potentially for inclusion in their research efforts. In addition to expanding the universe of targeted questions that can be answered, the rich set variables increasingly available for research can facilitate more exploratory efforts, evaluating differences, learning from data, and prediction. It is important to keep in mind, however, that these variables are captured in a much less structured manner and often without research efforts in mind (the challenges associated with this are discussed in a later section).

## Big "*t*"

Finally, the "always listening" nature of large digital platforms and instruments provides dramatic shifts in the ability to observe individuals and their behavior over extended periods of time and at a very granular level. Whether the engagement with the platform occurs today, tomorrow, or a month from now, the continuous nature of data collection allows for this engagement to be captured at reasonably low cost compared with traditional methods of data capture (although start-up costs of establishing the platform may be high). More so, if individuals engage with the platform, the platform captures changes in their behavior over very small time intervals (near-continuous time). Consider, as an example, the data generated by popular health wearables (e.g., Fitbit armbands) which typically capture minute by minute observation of step counts when worn (yielding missing data when being charged or not worn, itself creating interesting methodological issues that researchers may deal with in different ways). The ability to observe the behavior on these platforms in a semicontinuous fashion over prolonged periods of time allows researchers to precisely capture when events and behaviors of interest occurred, view these behaviors over long periods of time at a low cost (i.e., study long-term effects), and capture fine variation in these behaviors over these time periods. The ability to capture a rich set of variables fairly continuously could facilitate process-focused studies as well as "deep dives" into a single individual—reminiscent of what qualitative researchers argue

has been missing from quantitative approaches to studying psychology.

## General Challenges and (Some) Solutions

Although we believe that the changes in data available for research offer considerable benefits, they also come with notable and novel challenges. In this section, we introduce these challenges and some of their potential solutions. Because these challenges apply to a variety of big data research efforts, we introduce them in a general sense in this section. When we discuss specific opportunities for psychology in the following section, we delve into the instantiations of these challenges and how they can emerge differentially for different types of big data research effort.

### Getting Access

The platforms creating data that are "big" with respect to *n*, *v*, or *t* are diverse and growing (e.g., Amazon, Facebook, Twitter, Fitbit, Khan Academy). Many of these platforms involve large swaths of the (online) population while also capturing outcomes of interest to psychology. Of course, obtaining access to these platforms and the data they generate can be difficult, if even possible at all for researchers outside of the companies. For many academic researchers, this can seem like an insurmountable obstacle, and for good reason. Commitments from platform owners may be very difficult to obtain for a variety of reasons. Such commitments can be costly to platform owners; for instance, platform owners may have to expend resources to provide researchers access to their users and technology platforms (e.g., staff time). In addition to these direct costs, research collaborations may expose the organization to risk from negative press, disclosure of competitively relevant insights, invasion of customers' privacy, and actual (or sometimes perceived) violations of the terms of service of the platform.

Although access to large commercial data is not often within reach of academic researchers, we argue that this is not as significant a constraint as it is sometimes viewed, and that big data are increasingly within reach for researchers in psychology. This is because partnership with the platform owners may only be necessary if research efforts require that changes are made (e.g., introducing randomized treatment) for large swaths of the platform's users (i.e., big *n*) and/or if they require data on nonpublic interactions on the platforms (such as user logs or private interactions between users). If research efforts do not have these requirements, there are often alternatives to a direct partnership for accessing data. For example, researchers may gain access to some of these data through automated procedures that "scrape" data from public sites (e.g., Reddit, comments from a news organization, certain Facebook pages, etc.), or by purchasing it from third-party vendors (e.g., millions of user tweets can be purchased from multiple vendors). If research efforts do not require big *n*, users of these digital platforms (e.g., Facebook, Fitbit) can be directly recruited for research studies after which they can provide researchers permission to access the rich data that these platforms collect about them. These data can then be accessed through standard data requests to a platform's Application Program Interface (API).[1] Again, it is important to note that these approaches still have their own set of hurdles. For example, the approach of directly recruiting users from these platforms may become prohibitively costly if research goals require data on a large number (e.g., tens of thousands) of users, and these approaches require technical know-how (these challenges are discussed in more depth below).

There may also be opportunities to completely side-step proprietary digital platforms while still obtaining data similar to what the digital platforms offer. For instance, researchers may increase their sample size (big *n*) by leveraging digital crowd-sourcing platforms to solicit participants for research. These platforms are increasingly being used by researchers to efficiently and cheaply recruit study participants for studies designed and run by the researchers themselves (e.g., online surveys or experiments). In the crowd-sourced context, a widely used implementation is Amazon Mechanical Turk (AMT), which has been described as "the internet's hidden science factory" (Marder, 2015, p. 1); other similar platforms are gaining momentum and offer a comparable degree of data quality and efficiency (Peer, Samat, Brandimarte, & Acquisti, 2015). While it is not yet feasible to collect vast samples (e.g., those in the hundreds of thousands) via these platforms, they offer the potential to easily expand sample sizes via scalable computing architecture to several thousand individuals. Psychology, in fact, has already made progress in this context, with pioneering work validating AMT samples (e.g., Buhrmester, Kwang, & Gosling, 2011). There are also some potential limitations to crowd-sourced samples. For example, crowd-sourcing platforms can suffer from the emergence of power users, which results in a small portion of users accounting for a disproportionate amount of the activity (Paolacci & Chandler, 2014).

Diverse variables about individual behavior (big *v*) which are measured granularly over time (big *t*) may also be possible without direct access to these proprietary digital platforms. In particular, widely available research tools can provide rich data collection capabilities for researchers (i.e., in contexts that are or can be instrumented). For example, traditional survey tools commonly used to build questionnaires are becoming increasingly viable tools for building rich data collection environments. The Qualtrics survey tool, for example, has built-in questions that allow individ-

[1] An API is an interface often used by third-party developers to build software for and otherwise interface with a digital platform. This can be used to access data for users on these platforms.

ual participants to flow through survey information similar to how users would traverse an online website. Alongside these questions, Qualtrics also includes features that allow researchers to collect data on respondents' behavior in these environments. For example, Qualtrics allows researchers to time how long it takes to answer questions, capture the number of clicks and time spent on a page, and produce heat maps on pages to easily allow participants to indicate which sections of a page are most salient to them (down to the individual pixel on a screen).

The ability to collect data about individuals is expanded when using custom scripts and code that records outcomes of interest. For instance, Qualtrics supports custom JavaScript, which can be used to capture, store, and analyze detailed mouse movement data on the Qualtrics survey tool and capture the position of the mouse on the screen at a given point in time. Boas and Hidalgo (2013) integrate Qualtrics with rApache (a version of R that runs on Apache web servers), allowing them to dynamically generate content from outside sources (e.g., databases online) and perform analyses on the survey in real time. In some sense, it has become a misnomer to refer to such tools as "survey" tools when, in fact, they are tools for building what we describe as an interactive online environment that provides an instrumented way to collect rich data about the respondents and their actions. It is interesting that the burden of collecting data from many participants is easily scalable after the coding is complete, unlike many studies in which one or more researchers are involved one-on-one in data collection efforts (e.g., visits to the laboratory).

In addition to the potential of advances in survey tools to facilitate the development of an interactive online environment that allows for richer data collection, researchers have begun to develop custom packages that enable researchers to conduct elaborate, natural, and instrumented web experiments. For example, de Leeuw (2015) provides an open-source javascript library (jsPsych) that can be integrated into a website to provide rich data collection capabilities. Garaizar and Reips (2014) developed an open-source Web based system that simulates a social networking environment and captures data on how participants navigate and communicate on the platform. These frameworks, which are openly available without the partnership of large platform holders, present a number of possibilities for collecting data that have not thus far been used in psychology but that can provide rich insight in a wide variety of contexts.

Coupling these varying and highly accessible options, it is easy to imagine a study in which researchers solicit thousands of participants in the span of a couple of days to take part in experiments or observational studies using naturalistic research environments that are quick to develop and capable of granular data collection as well as advanced logic and functionality. Although these data may not have some aspects of the data collected from a proprietary digital platform (e.g., the realism of the decision context), the benefit is that data collection is accessible and under the direct control of the researcher and can often be conducted at reasonably low cost.

## Technical Challenges

Even after gaining access to big data, researchers will likely find that the rich set of variables captured granularly over time by these platforms often requires considerable processing and cleaning before becoming useful for research efforts; this harkens back to Coombs (1964) and the importance of parsing collected observations into meaningful data. These challenges emerge primarily because the rich set of variables (big *v*) captured by these digital platforms does not come in a neat format easily incorporated into research efforts. Moreover, data collected over time from these platforms can be entirely different depending on the measurement occasion; moreover, such data can be measured at different levels of granularity, do not necessarily come from the same sources, and in fact might be open-ended responses. This variety in the types of observations speaks to the difficulty of parsing observations into data before use. In fact, scholars have commented that in the context of big data analysis, 80% of the time is spent preparing data and only 20% on analysis (e.g., Wickhan, 2014; see also Dasu & Johnson, 2003) and that "it's an absolute myth that you can send an algorithm over raw data and have insights pop up" (Lohr, 2014, p. 2). To illustrate the unique type of noise and data cleaning needed on digital platforms, consider that a recent study suggests that nearly 40 million Twitter accounts are actually automated bots designed to mimic user behavior online, this study also offers a classification framework for identifying and accounting for these fake accounts (Varol, Ferrara, Davis, Menczer, & Flammini, 2017). This supports the more general point that research conducted with these data may require a nontrivial degree of technical expertise to simply administer, manage, and to "wrangle" and "tidy" the data; the required technical expertise might be even more pronounced if the effort involves randomized manipulations and data collection from live and dynamic platforms. These concerns are exacerbated by the scale of individuals (big *n*) and the speed at which data on them accumulate (big *t*) since any manual approach toward data cleaning (e.g., research assistants or manual coders) quickly becomes infeasible, necessitating the use of automated scripts and coding to clean and process available data.

Recent efforts by scholars in psychology and related fields are starting to address the technical challenges of accessing and processing big data. For example, Chen and Wojcik (2016) offer a practical guide to identifying big data sources for research, approaches for collecting data, and methods for processing and analyzing data. Landers, Brusso, Cavanaugh, and Collmus (2016) offer guidance on

automated extraction of online data through web scraping. There are also specific software packages targeted toward specific digital platforms: twitteR (Gentry, Gentry, RSQLite, & Artistic, 2016) and RedditExtractoR (Rivera, 2015) offer toolsets for extraction of data from Twitter and Reddit, respectively. Dehghani et al. (2016) created the Text Analysis, Crawling, and Interpretation Tool (TACIT) as an open and extensible tool coupling capabilities that allow for the collection and analysis of large-scale text data. Leaning on existing software packages may be an important way for traditional researchers to leverage known skills while also dealing with other technical and statistical challenges of big data research. Chen and Wojcik (2016) note that "although computing skills are necessary for big data research, expert-level abilities are generally not required, in part because of the availability of preexisting software libraries that implement advanced techniques" (p. 459).

## Making Sense of Big Data

Simply because a digital platform offers access to large numbers of study participants, and rich and diverse data about them, does not mean that these available data are immediately useful for studying topics of interest to psychology or that they will extend the literature. While scholars (Jaffe, 2014) have posited that the "data breadcrumbs" left online via clickstream data may reveal fundamental individual characteristics (e.g., personality, cognitive style, emotion), actually translating these "data breadcrumbs" (which were not collected with research in mind) into constructs and outcomes of interest to psychology is not trivial and presents important challenges for psychological researchers. In particular, well-validated survey instruments that measure psychological features and constructs (e.g., emotion, personality, intellectual ability, etc.) are rarely administered to users of these large digital platforms, leaving a serious conundrum for researchers. Again, we find that this challenge is exacerbated by the scale of unique individuals (big *n*) available for study on these platforms because this often precludes researchers from using traditional methods (e.g., those from psychometrics) for measuring constructs and evaluating relationships among constructs.

Whereas the specific methods used will vary, we suggest that the problem is conceptually similar to linking or equating tests, an idea widely used in educational measurement contexts (e.g., Dorans, Pommerich, & Holland, 2007). Extending linking and equating ideas to a big data research setting suggests that one way to "make sense" of big data is to attempt to link or equate two sets of variables: the highly validated measurement instrument (traditional) and raw data captured by these instrumented platforms. One practical approach for doing this is to measure constructs of interest (using traditional methods) for a subset of users on a platform of interest, then leverage the availability of both structured and raw data for this subset of users in an effort to start to form models that help equate these raw data to constructs of interest. This could be an application of a planned missingness design, albeit a more extreme version (e.g., Rhemtulla & Little, 2012; Silvia, Kwapil, Walsh, & Myin-Germeys, 2014), where intensive measures are administered to a small subsample of a larger dataset.

It is difficult to overstate the importance of efforts that help translate the variables available on big data platforms to constructs and measures of interest to psychology. First, linking a rich set of variables (big *v*) that accumulate granularly over time (big *t*) to constructs of interest to psychology may reveal important features of the constructs themselves, as well as how these constructs evolve over time. Second, and maybe even more important, by building models from a subset of individuals, we can start to discern psychological constructs from observed data available for many more individuals and hopefully start to evaluate the relationship between these constructs and behaviors at the scale of individuals that these big data platforms provide. Consider as a case in point that the Cambridge Psychometric Centre has recently created an API (which is accessible for free to researchers) that leverages models (learned from linking Facebook data to psychological constructs) to generate predictions of personality, happiness, intelligence, and so forth based on Facebook likes and messages. Such tools open up a vast set of research questions for a wide range of researchers by allowing any researcher with Facebook data to consider questions related to constructs for which they aren't able to administer the surveys typically used to measure them.

## Statistical Challenges

The scale of individuals, breadth of variables, and granularity at which they are collected also introduce the challenge that traditional approaches for statistical analysis as well as the interpretation of results may no longer fit well. Similar to other challenges, the various dimensions of big data introduce different types of statistical challenge.

With increased sample size or "big *n*," relationships in data will tend to be highly statistically significant, necessitating a discussion of the magnitude of effects and their importance. Confidence intervals for the population effects, for example, may be narrow and not include zero (i.e., be statistically significant) but bracket values that are of a size such that little value is obtained. That is, the full set of values in the confidence intervals may not be beyond the "good enough range" to consider them of any theoretical importance (i.e., the size of the true effect is at best close enough to the null value for it not to be theoretically interesting; e.g., Serlin & Lapsley, 1993. If data also include a rich set of variables about individuals (big *v*), potential for spurious correlation makes any interpretation of *p* values for

any single variable even more problematic. e.g., attempting to correct for multiplicity issues when many hypotheses are tested is an ongoing topic in behavioral genetics, in which many genetic variants are evaluated for potential explanatory value (Troendle & Mills, 2011).

This combination of factors suggests that, with these data, an alternative focus may be on achieving high predictive validity through rich statistical models coupled with cross-validation to ensure good out-of-sample prediction (Domingos, 2012). The challenge with this, however, is that many commonly used statistical approaches in psychology (and many other fields) are limited in their capability to handle complex models that include a large number of predictors (e.g., Hastie, Tibshirani, Friedman, & Franklin, 2005). The limitations of traditional methods to handle this increasingly rich set of variables have given rise to methods developed at the intersection of statistics and computer science, such as machine learning, which is especially applicable in situations of large or rich data sets (e.g., Domingos, 2012). Historically, machine learning approaches have tended to be data-centric, loosely guided by theory if theory is used at all, with a focus on classification, pattern recognition, and prediction. However, as machine learning approaches become more broadly used, there has been a rise in methods that can be used in a more theory-driven fashion. For instance, advances in methods for topic modeling allow researchers to focus on specific topics or areas of interest in unstructured data (Andrzejewski, Zhu, & Craven, 2009; Wang & Blei, 2011). Brandmaier, Prindle, McArdle, and Lindenberger (2016) propose a method that joins structural equation modeling and decision trees to allow for automated selection of variables that predict differences across individuals in specific theoretical models. Other approaches focus on identifying heterogeneous treatment effects in secondary and experimental data (e.g., McFowland, Speakman, & Neill, 2013) and are likely to be highly conducive to theory building and theory generation.

At the same time, some well-known statistical issues are exacerbated (relative to traditional efforts) by machine learning used in conjunction with big data. For example, concerns of overfitting and the "curse of dimensionality" emerge as some of the most pressing concerns associated with these approaches (Domingos, 2012). This can be particularly true when data include many variables collected for relative few individuals, resulting in, for example, artificial increases in least squares model fit (James, Witten, Hastie, & Tibshirani, 2013). Moreover, additional challenges arise if variables are collected in near-continuous time (big *t*) because such variables are not likely to be measured at neatly structured occasions, there may be missing or "not-applicable" data, and synchronicity often will not hold. In these cases, researchers may not only need to consider models that can handle a rich set of variables available for research, but also those that can accommodate time-series data being rapidly created and included in analysis (Ding, Trajcevski, Scheuermann, Wang, & Keogh, 2008; Xi, Keogh, Shelton, Wei, & Ratanamahatana, 2006).

Again, recent efforts are starting to help scholars in psychology overcome the challenges with analyzing these types of data. For instance, Stanton (2013) offers a practical introduction to data mining efforts in psychology with details on the steps necessary to transform raw data into (usable) processed data so that statistical models and analyses can be implemented and interpreted. Other works offer an introduction to machine learning methods with a focus on applications in psychology and related fields (Oswald & Putka, 2015).

## Theoretical and Research Value

Highly unstructured and varied data without clear measures of interest to psychology coupled with the use of statistical approaches that are often viewed as atheoretical introduces the challenge that the efforts enabled by big data may be useful in applied problems, but could be limited in their value to develop, inform, or evaluate psychological theory of underlying phenomena (e.g., because interpretation of individual predictors is not the focus). This concern reinvigorates what is actually a longstanding discourse on the role of exploratory or predictive efforts relative to explanatory ones (e.g., see Pedhazur, 1997).

We argue that these concerns may be warranted only to a point; the opportunities that leverage these rich data may take several forms, many of which may have direct and considerable promise to inform theory (although perhaps in an inductive fashion). In particular, it is partially a misconception that big data research need be pursued in an atheoretical fashion; as we will see in the following section, research efforts that leverage diverse forms of big data employ theory-driven approaches. This includes research that is driven by clear theoretical tensions in the literature, that collects and aggregates data with the explicit purpose of testing ex ante motivated and stated hypotheses, and then tests these relationships. That said, these studies do often need to apply novel statistical approaches to these data which allow them to extract constructs and measures of relevance to the theoretical frameworks and questions of interest to them.

Moreover, there may be significant research value in making accurate predictions and exploratory analysis if they allow us to better understand and potentially change behavior. This has clear links to efforts seeking to encourage behavioral changes for positive outcomes, something that many areas of psychology care about deeply. Interventions to encourage various behaviors could even be personalized based on what a participant's data reveal about his or her psychological features or dispositions. These ideas are akin to personalized medicine for psychological outcomes. In addition, prediction or exploratory efforts have the potential

to reveal variables that do not necessarily seem theoretically grounded in those phenomena. For instance, learning psychological constructs from unstructured data may reveal important nuances about the constructs themselves, which could guide additional theoretical development and evaluation. In other words, what initially does not seem theoretically grounded may simply not yet be incorporated into a theory, and theories may be developed based on findings from exploratory or predictive big data research. Perhaps the instrumented world is exactly what is needed to more rigorously test and validate psychological theories and to turn research findings into practical value. This may necessitate, however, the field becoming more open to insights originating from less theoretically rigid starting points.

## Nuanced and Varied Opportunities for Psychology

The intersection of the two previous sections, in which we discuss the various benefits and challenges associated with big data, results in nuanced and varied research opportunities for psychology. These opportunities may first be considered with respect to only a single dimension of big data. For instance, researchers could expand their study sample (big *n*) but retain their focus on a small number of variables (little *v*), measured cross-sectionally (little *t*). This approach presents an opportunity to address underpowered studies in psychology and, if samples are more representative, reduce concerns that results may not generalize to the population of interest. Similarly, the potential to collect a richer set of variables for analysis provides the opportunity to evaluate more targeted relationships and the interactions (moderators) of relationships that are of potential interest. The ability to collect these variables more granularly over time may alleviate concerns of dynamic constructs that have been historically measured too coarsely (e.g., only morning and evening blood glucose levels, whereas blood glucose levels can now be measured near continuously when assessing, e.g., mood or distractibility). While these opportunities are nontrivial, they may not fully leverage the benefits provided by big data. For example, statistically significant effects need not imply that effects are of practical significance (Kirk, 1996; Kelley & Preacher, 2012), and gains in statistical power and estimation precision for few parameter estimates are less relevant as sample size increases beyond a certain point; effect sizes of practical significance typically do not require hundreds of thousands of individuals to be precisely identified (e.g., obtain a narrow confidence interval). Similarly, as we noted, using a rich set of variables for traditional research efforts (i.e., highly targeted) may exacerbate concerns about spurious correlation and selective reporting of significant effects in data.

We contend then that important opportunities for big data research in psychology emerge at the intersection of growth in these various dimensions of data; this is because the benefits of big *n*, *v*, and *t* are highly complementary to one another. Consider for instance that the ability to collect data over extended periods of time (big *t*) can be critical to fully leveraging the large numbers of individuals available on these platforms for research (big *n*). This is because large-scale adoption of these platforms does not equate to continuous use and activity by platform users. For example, some sources suggest that only 44% of Twitter's users have ever sent a tweet (Koh, 2014). Thus, extended windows of observation (historically and in some cases prospectively) are critical to actually having data on a large swath of users from these platforms (over short periods of time, there would be sparse or no data for many of these users). In such cases it is not that data are necessarily missing, it is simply that the individuals did not utilize the service and thus no data are available. Moreover, exploratory or predictive efforts that fully leverage a rich set of variables (i.e., big *v*) require considerable variation in these variables if they are to be feasible. This suggests that a rich set of variables (big *v*) may need to be coupled with a large sample (big *n*) or measurement that is rich longitudinally (big *t*) to unlock their full potential. C. R. Rao suggests that big samples enable new types of research efforts because "certain assumptions of theoretical models can be relaxed, over-fitting of predictive models to training data can be avoided, noisy data can be effectively dealt with, and models can be validated with ample test data" (Nielsen, 2016, p. 4). In addition, Oswald and Putka (2015) highlight that while "one might think that datasets that one might call big" would alleviate concerns of underpowered studies, "big data models are able to fit complex relationships where they exist, and therefore they too are very hungry for data" (p. 45). Relatedly, the fact that for various combinations of variables in big data, the number of observations (e.g., in a particle "cell" or combinations of variables) can be very sparse is often lost amid the excitement that surrounds big data.

In the rest of this section we offer an organization of research opportunities for psychology by how they leverage some combination of large sample size (big *n*), a rich set of variables about individuals (big *v*), and granular and sustained data collection over time (big *t*). We provide citations for work that has had to deal with difficulties, not necessarily because they did so in either the ideal way or via a poor approach; they simply serve as exemplars.

## Big *n*, Little *v*, Big *t*: Or, Traditional Research Expanded

We first consider research efforts focused on targeted questions (i.e., few variables) while still leveraging a combination of big *n* and big *t* to capture these interactions in a real-world context and for many individuals. This variant of big data research has the potential to inform inquiries about

a specific set of variables and their relationships to one another while also observing actual behavior in the real world. These efforts may be highly effective for research areas in which theories are established but have persistent tensions; particularly if these tensions stem from underpowered studies, sample bias, or concerns of the realism of the research contexts in which they are studied.

Some recent studies exemplify this form of big data. For example, Muchnik et al. (2013) conducted a randomized field experiment over a 5-month period and found evidence of bias caused by social influence on a popular social news site. Leveraging the large user base of the site (big *n*) in combination with the ability of the site to continuously collect data over time (big *t*), they manipulated 101,281 comments that were then viewed more than 10 million times and rated 308,515 times over the 5-month study period. Because they were able to observe behavior over extended periods, they were also able to observe that positive social influence had accumulating herding effects that increased its effect over time. Another example is the work of Bapna, Ramaprasad, Shmueli, and Umyarov (2016) who, over a 3-month period, conducted a large-scale field experiment on a popular online dating website. They manipulated whether 100,000 users on this dating site were provided a popular privacy feature that lets them hide their perusal of others' profiles; they found that women received significantly fewer romantic interactions on the website if provided this privacy feature. They conjectured that privacy features were reducing romantic interactions for women because women were benefiting from an indirect mechanism for signaling romantic interest. Similar to Muchnik et al. (2013), the combination of a large user base on the platform and the ability to collect data over time was key to their ability to validate and quantify their results (e.g., they were able to observe pretreatment trends as well as posttreatment effects). Similarly, Kramer, Guillory, and Hancock (2014) uncovered causal evidence of emotional contagion by randomly presenting nearly 700,000 Facebook users' content with varying degrees of emotional valence ($n$ = 689,003) over a 1-week period. Exemplifying the role of time, only those who posted a status during the week of the experiment ($n \sim 465{,}000$) were included in the study.

For these types of efforts, challenges related to access are notable. For example, randomized field experiments on a large digital platform are likely to require significant and difficult-to-secure commitments from platform owners (functionally making them often inaccessible to researchers). All of the studies discussed in this section leveraged some type of relationship with platform owners as they required the ability to introduce controlled and randomized manipulations into these environments for a large number of users. More so, the ability to collect detailed data on individual behavior was often key to their research questions. For instance, Bapna et al. (2016) identified their core outcome of interest by evaluating otherwise private messages sent between users of the platform. Moreover, their intriguing insight of gender asymmetry in the initial steps of the dating process relies on microlevel data captured granularly over time (i.e., data on who viewed someone's profile and whether messages followed these views or not). Of course, research efforts of this type need not be randomized field experiments (e.g., secondary data sets can also leverage few outcome variables and few explanatory variables). However, the use of secondary big data to evaluate targeted questions suffers limitations similar to those affecting more traditional efforts, including limitations related to potential omitted variable bias and measurement error.

Another consideration with these types of big data efforts is that the scale of individuals involved often precludes additional data collection by researchers, resulting in measures that may approximate outcomes or variables of interest. For instance, Kramer et al. (2014) analyzed Facebook posts for emotional valence and categorized posts as "positive or negative if they contained at least one positive or negative word" (p. 8789). In addition, Bapna et al. (2016) utilized an exchange of three messages as a proxy for whether a match occurred on the platform, but could not observe whether individuals actually went on a date, the quality of that date, or whether participant matches resulted in a relationship. More so, there are technical challenges related to the administration of these studies. For instance, Kramer et al. (2014) had to modify software that analyzes text for positive versus negative words to run on Hadoop (a software framework designed for applications that can process massive amounts of data) and in participants' Facebook news feed. Finally, there are experimental design considerations associated with research on these platforms, such as introducing manipulations at rates equal to their natural occurrence on the platform. Because of this concern, out of 101,281 comments included in the experiment conducted by Muchnik et al. (2013) only 4,049 were positively treated (assigned an up-vote) and only 1,942 were negatively treated (assigned a down-vote). Similarly, Kramer et al. (2014) ran two separate control conditions for positive and negative valence treatments because posts with a positive valence were more common on Facebook valence.

On the other hand, some challenges that can emerge in big data research may be less pronounced. First, statistical analysis for these studies may be reasonable straightforward, as such studies often use standard estimation approaches. Bapna et al. (2016) rely primarily on evaluating differences in means and *t* tests, Kramer et al. (2014) utilize Poisson regression and weighted linear regression, and Muchnik et al. (2013) use more complex (but still well known) regression models that include random effects to account for repeat observations from the same rater. That said, the large sample size means the effects identified by these studies tend to be statistically significant, necessitating a dis-

cussion of whether the effects identified are of theoretical or practical importance. Second, these studies tend to be highly focused and may be easier to position in terms of their theoretical value and research; in fact, all of the studies in this subsection are motivated by fairly specific limitations in prior works (e.g., external validity or causal interpretation of prior results). The applicability of existing statistical approaches and the potential of these studies to be theory-based thus allow this type of data and research to be readily integrated into existing research frameworks. At the same time, such efforts can also be considered a figurative "dip of the toe" in the grand scheme of what is made possible by big data.

## Little *n*, Big *v*, Big *t*: Or, Small Sample Big Data Research

Another form of big data research leverages an increase in the variables available for researchers as well as granular changes in these variables over time, rather than focusing on the sample size (*n*), per se. This type of big data research has garnered the least attention from a big data perspective and yet, in our opinion, offers some of the most actionable potential to make immediate and tangible contributions for researchers in psychology to contribute to the literature. These efforts can yield important process-oriented insights, and can also include exploratory or predictive research that seeks to link traditional psychological constructs to the types of variables typically generated by big data platforms.

A number of studies exemplify what is possible with small sample big data research. Particularly promising are research efforts that combine both traditional measures of interest in psychology and rich granular data on individuals' behavior. Wang et al. (2014) collected data on 48 students over a 10-week spring term which included 53 GB of sensing data from their smartphones (e.g., location information, sleep, activity information, duration and frequency of face-to-face conversations, etc.), 32,000 daily self-reports covering variables such as affect, stress, exercise, mood, and loneliness, pre and post surveys measuring psychological constructs of interest (personality, depression, etc.), and measures of academic performance (e.g., GPA). Using these data, Wang, Harari, Hao, Zhou, and Campbell (2015) were able to predict GPA within +/0.179 of the reported grades and identify a number of important determinants of academic performance. Using the same sample, they show that data commonly recorded by smartphones can strongly correlate to changes in daily stress level, depression, and loneliness (Ben-Zeev, Scherer, Wang, Xie, & Campbell, 2015). Purta et al. (2016) captured rich Fitbit data (e.g., steps, calories burned, sleep, etc.) and cell phone data (sensor data, phone logs, etc.) for 500 undergraduate students over a 2-year period. Concurrently, they periodically captured survey data on personality, self-esteem, self-efficacy, mental health, overall health, sleep habits, physical activities, exercise, and affiliations. Mark, Iqbal, Czerwinski, Johns, and Sano (2016) automatically logged computer activity at work during all business hours over 12 work days. They also administered a general survey measuring the Big 5 personality traits as well as impulsivity, stress, and other variables. They leveraged this data to better understand distraction, focus, and productivity at work; for example, they found that neuroticism was associated with shorter online focus duration and that some individuals were more susceptible than others to online attention shifting in the workplace. Hibbeln, Jenkins, Schneider, Valacich, and Weinmann (2017) leveraged three small sample studies ($n$ = 65, 126, and 80) and an experimental approach merged with granular capture of mouse movements to evaluate the relationship between mouse cursor movements and negative emotion. They found across the three studies that cursor distance and speed were significant predictors of negative emotion (in one study they were able to identify groups who were provided the negative emotion treatment with 82% accuracy). In each of these research efforts, the data sets were "big" and varied but dealt with a relatively small number of individuals.

A few challenges emerge as most prominent with this type of research effort. First, there is often a significant need to preprocess and then analyze raw data to generate variables that are useful for the work at hand. For instance, Wang et al. (2014) created scripts to automate data collection from the smartphone's accelerometer, light sensor, microphone, and gps/Bluetooth, and then employed a variety of classifiers to convert this raw data into variables of interest. In particular, they extracted features from preprocessed accelerometer data and applied a decision tree classifier to infer physical activity, used a separate set of classifiers to infer human voice and conversation from microphone data, and built a classifier that uses data from both a phone's light sensor and microphone to build a classifier that discerns when participants are likely asleep. Similarly, Wang et al. (2015) created approximate measures of social versus academic behavior using location information (are they in a library or a fraternity), the level of ambient noise, how long they stay in a particular location, and so forth However, researchers may not always need to perform this type of analysis themselves; for instance, Purta et al. (2016) leveraged Fitbit's algorithms to approximate participants' levels of physical activity, sleep patterns, and so on. Of course, had others implemented their own algorithms the answers could be different. These sorts of measurement problems that psychologists have long considered (reliability and validity) should be, in our view, more front-and-center in big data research.

A second notable challenge related to these efforts revolves around the statistical analysis of a rich set of variables collected longitudinally for relatively few unique individuals. Wang et al. (2015) note that the small number of unique participants in their sample introduced overfitting

concerns and limited their ability to use sophisticated predictive models. Ben-Zeev et al. (2015) faced similar limitations and used penalized functional regression (Goldsmith, Bobb, Crainiceanu, Caffo, & Reich, 2011) in order to "use intensive repeated-measure variables as predictors and relate them, as a whole, to individual-level outcome measures" (p. 6). Ben-Zeev et al. (2015) also accounted for the rich longitudinal nature of data in these samples by employing mixed effects linear models. A third challenge relates to ongoing compliance by participants, particularly for studies that run over longer periods. In particular, data collected from digital devices (such as smartphones or a Fitbit) require that these devices be charged, that they be carried by participants or worn on their person, and that certain functionality (e.g., Bluetooth) be consistently kept on. This requires researchers to implement ongoing incentive schemes to maintain high compliance by participants throughout the study period. For example, Wang et al. (2014) raffled prizes throughout the study period to high compliance participants, and Purta et al. (2016) provided a monthly cash stipend for participants that was dependent on their compliance rates.

Other challenges associated with big data research are less pronounced with the efforts we are considering in this section. First, access may be less of a challenge because these efforts do not involve commercial interests or platforms; all of the studies discussed in this section were conducted without collaboration with technology platform owners and are primary data collection efforts by researchers. Hibbeln et al. (2017) illustrate how these efforts may be conducted using only traditional samples, open source tools, and researchers' own experimental environment. Specifically, they utilize one sample from Amazon Mechanical Turk and two student samples in conjunction with digital environments they developed themselves which granularly capture user mouse movements. They were able to develop webpages captured the mouse cursor's position and timestamp at a millisecond precision rate by using a publicly available JavaScript library (JQuery). Through this approach, they were able to introduce randomized manipulations into these environments and also efficiently examine their hypotheses across different decision settings. The accessibility of this form of big data research and its ease of replication can offset some of the selection and generalizability concerns associated with the small samples. This is because other researchers can more easily replicate these studies to validate, complement, refute, or find and evaluate moderators or mediators for prior results. Second, challenges related to theoretical and research contribution may be less pronounced since researchers have more direct control over data collection and research procedure. This can lend to research efforts that employ rich data from digital platforms to test preplanned and theoretically driven hypotheses. For example, when collecting granular and rich data about individuals' digital activity, Mark et al. (2016) formed and tested specific hypotheses about how this digital behavior relates to psychological constructs. Relatedly, the small sample allows researchers to use traditional approaches to measure constructs of direct relevance to psychology (e.g., survey approaches). In fact, these types of effort may be critical for "making sense" of big data for psychology by linking traditional psychological constructs to unstructured but rich variables generated by these platforms.

## Big *n*, Big *v*, Little *t*: Or, Small Snapshot Big Data Research

Also holding promise for psychological research are big data research efforts that leverage a rich set of variables in conjunction with large samples, while not necessarily relying critically on measurement of variables and observation over time. In these studies, large samples can be key to unlocking the full potential of a rich set of variables to inform questions of interest to psychology and to facilitate exploratory and predictive efforts.

One area in which this type of research effort is relevant and can provide novel exploratory and predictive results is behavioral genetics. In much of this research there are many single-nucleotide polymorphisms (SNPs) recorded (big *v*); indeed, there are often many more SNPs within an individual that are of potential interest to researchers than there are individuals in these studies. Thus, the ability to observe many individuals (big *n*) is key to uncovering robust relationships between this genetic information and outcomes of interest, particularly because genes do not change, though gene expression (i.e., epigenetics) and outcomes of interest can. For example, Hu et al. (2016) leveraged data on 89,283 individuals to identify genetic variants associated with self-report of "being a morning person," and how being a morning person relates to a various psychological outcomes including insomnia and depression. Another example of big data research that may not require granular variation over extended periods of time is modern-day text analysis, in which rich text captured over a short period of time can generate a wide variety of variables and features and can be available for a large number of individuals. For instance, Dehghani et al. (2016) collected 731,332 tweets from 220,251 users (188,467 of which they were able to collect network structure for) in order to evaluate how dimensions of morality predict social distance online; they hypothesize and find confirming evidence that "moral purity" exceeds other dimensions of morality in terms of explaining social distance.

Researchers considering these types of efforts should again be cognizant of some of their most prominent challenges. For instance, the combination of variable richness and large sample sizes raises considerable concerns around spurious correlation and effect sizes that are statistically significant but not practically (or clinically) significant. These concerns are clear in the context of behavioral genetics where genome-wide association

studies evaluate the relationship between millions of SNPs and sometimes few outcomes. Because of this concern, research efforts lean on revised norms for statistical significance, using $p < 5.0 \times 10^{-8}$ as an acceptable threshold for what is considered a statistically interesting association (Hu et al., 2016).

In addition, the scale of the participants in these studies makes it difficult to collect data on constructs of direct relevance to psychology. Some studies side-step this issue by leveraging proprietary data that join a rich set of variables with outcomes of interest to psychology. For example, Hu et al. (2016) were able to conduct their study looking at genetic indicators of being a morning person (as well as related outcomes like depression) by leveraging self-reported survey data collected by "23 and Me" (a large-scale but proprietary digital platform that sells low-cost genetic mapping kits). More common with these research efforts, however, is that constructs of interest are not directly observable in available data (particularly when the scale of individuals is large). For example, Dehghani et al. (2016) were interested in the impact of morality on social distance but could not use traditional survey measures to capture moral leanings in their sample. Instead, they "capture morality more indirectly by observing the naturally occurring 'moral residue' left behind in the texts of social discourse" (p. 2). They do so by using singular value decomposition to reduce tweets to vectors that can be compared in terms of their distance to terms associated with a particular moral concern; this then allows them to evaluate how different moral concerns relate to social distance online. Illustrating the importance of efforts that link observed data to constructs of interest to psychology, the approach developed by Dehghani et al. (2016) could be used by other researchers to evaluate a wide variety of other questions related to moral leaning for a potentially large swath of the online population.

A final consideration for researchers interested in these kinds of efforts is that while the use of secondary and large-scale data can reveal interesting insights, the lack of researcher control they introduce as well as the need to approximate constructs of interest can make it difficult to draw robust causal conclusions. Hu et al. (2016) addressed this issue through Mendelian randomization analysis and actually found that the initially strong correlations between being a morning person and psychological outcomes (e.g., depression) were not robust to more stringent causal analysis. Dehghani et al. (2016) addressed the same issue by complementing their large-scale observational study with two lab experiments using traditional measures of moral concern as well as random assignment to confirm their results.

The combination of large samples and rich data about individuals in these samples has the potential to be valuable for a range of research efforts in psychology. In particular, these efforts have significant potential to inform questions of interest to psychology that focus on variables that do not vary significantly over the time period examined, as well as those where small snapshots of many individuals provide sufficient richness of data to address research questions of interest.

## Big *n*, Big *v*, Big *t*: Or, Idealized Big Data Research

We now consider research efforts that intersect all three dimensions of big data with large samples and a rich set of variables that are captured granularly over extended periods of time. This type of study represents some of the more transformative potential of big data in psychological research. In other disciplines, such data have enabled research efforts that have garnered considerable public interest.

Aral and Walker (2012) first employed a large-scale digital experiment (1.3 million Facebook users over a 44-day period) to understand the factors that make individuals susceptible to social influence, clearly a construct of interest to many psychologists. Interestingly, the study leveraged the social network of only ~8,000 unique users of a mobile application to run an experiment on 1.3 million Facebook users; this highlights both the power of networked applications in enabling large-scale studies and the ethical concerns they can elicit (which are discussed in more depth below). The authors then extrapolate their experimental results to a larger secondary dataset (12 million users with 85 million connections) to find that influential individuals are less susceptible to influence than noninfluential individuals and that influencers tend to cluster networks while susceptible to influence do not.

Kosinski, Stillwell, and Graepel (2013) used a combination of historical Facebook data (e.g., "likes," detailed demographic profiles, etc.) from 58,000 volunteers and results of several psychometric tests to build predictive models that are able to infer some sensitive characteristics about individuals (e.g., sexual orientation, race, or political affiliation). Youyou et al. (2015) used similar data (from 90,000 participants) to evaluate the ability of computer models joined with data on user "likes" on Facebook to predict variation in individual personality characteristics. They found that data-driven models were more accurate than participants' Facebook friends in predicting personality measures, and that the computer personality judgments had high external validity when predicting life outcomes such as substance abuse, political attitudes, and physical health. Other efforts leverage centuries of text data to evaluate changing psychological states and constructs over extended periods of time. For example, Iliev, Hoover, Dehghani, and Axelrod (2016) analyzed two centuries of texts from historical corpora of American English—Google Books and the *New York Times*—to show that the well-documented linguistic positivity bias exhibits a linear time trend over time

and that this effect is sensitive to changes in economic, social, and psychological factors. Iliev and Smirnova (2016) undertook a similar effort and evaluated historical corpora to validate the hypothesis that causal cognition (as revealed by changes in causal language over time) has an increasing role in western society over time.

These efforts represent exciting work at the intersection of big data and psychology. It is unsurprising however that these efforts also face the full gamut of challenges discussed previously and thus may be furthest from the comfort zone of typical researchers in psychology. The analysis for these types of efforts can be complicated and require leveraging advanced statistical methods to both process available information to make it usable for analysis and then to analyze these data. For example, Iliev and Smirnova (2016) and Iliev et al. (2016) both utilized advanced text analysis methods to extract measures of interest to psychology (extent of positivity bias and casual cognition respectively). Kosinski, Stillwell, and Graepel (2013) first reduced the dimensionality of data on individuals' Facebook likes using singular-value decomposition. They then used top-100 SVD components in combination with cross-validated linear and logistic regression models to predict various outcomes of interest in the data. Youyou et al. (2015) used similar data but instead of reducing the matrix representing the participants' likes, they combined these data with LASSO regression (an estimation approach more adept at handling rich data for prediction, especially when some coefficients are close to zero and are thought to functionally be ignored). Similar to other works, they also used cross-validation to avoid overfitting of models, and repeatedly trained models on different subsets of data in order to generate predictions for their entire sample.

Time also plays a more significant role in the analysis and data collection for these efforts, both in terms of the theoretical novelty and value as well as their methodological approaches and challenges. For instance, Iliev and Smirnova (2016) and Iliev et al. (2016) proposed and tested explicit hypotheses related to variability in written texts over time and how these texts change in response to historical events and societal changes, again over time. In other instances, variability over time poses methodological challenges for such efforts. For instance, Kosinski, Stillwell, and Graepel (2013) achieved relatively low prediction accuracy for measures that are highly temporal in nature such as self-reported "satisfaction with life" and relationship status. They note that this low prediction accuracy may be because Facebook likes accrue over a longer period and may thus be best suited for predicting more long-term and stable measures. While these types of efforts may face many of the challenges associated with big data, it is important to note that they may be highly flexible in terms of research aims. For instance, although all of these efforts leverage data rich along all three dimensions of big data, they span the spectrum in terms of their level of ex ante theoretical development and testing: Iliev and Smirnova (2016) and Iliev et al. (2016) motivate and test specific hypotheses, whereas Aral and Walker (2012) start with a targeted experiment and then extend into more exploratory analysis, and Kosinski, Stillwell, and Graepel (2013) and Youyou et al. (2015) take exploratory approaches focused on predicting individual features.

These efforts may be the most valuable, at least in the short term, in areas in which prediction efforts offer clinical value, or where exploratory efforts have the potential to uncover insights into the relationship of psychological constructs with observed behavior. Long-term, these efforts could be valuable in areas where theory is underdeveloped or where existing theory fails to predict or explain robust empirical phenomena. As such, they may be highly useful for exploring and identifying novel theoretical directions.

## Pairing Different Research Efforts

Researchers need not restrict themselves to a single form of research because, in many cases, a single research article can be made stronger by simultaneously pursuing different forms of research (both traditional and leveraging big data). Simplest, but perhaps most powerful, is the pairing of big data research with traditional approaches used in psychology. In cases where scholars identify interesting correlations or potential insights but need to cement results through stricter causal analysis or more focused examination, traditional approaches can be a suitable complement. Moreover, researchers may have employed approximations of a construct of interest and may benefit from validating results with small samples and traditional methods for measuring the same construct (e.g., survey instruments). In addition, small sample big data research (little *n*, big *v*, big *t*) that tries to estimate how rich but unstructured variables correlate with more traditional psychological constructs can help larger-scale studies (e.g., big *n*, big *v*, and big *t*) "make sense" of their data and focus their exploratory or predictive efforts. Equally, these large-scale exploratory efforts could yield results that are suggestive of more fundamental relationships worthy of more targeted exploration. In this regard, randomized experiments on the same platform (e.g., through big *n*, little *v*, big *t* efforts) could be useful complements to these efforts because they can more cleanly capture and disentangle relationships between prespecified variables and do so in the same setting as the initial evaluation. Ultimately, how to pair both traditional and big data research is an issue of recent debate in the field (Baumeister, 2016; Sakaluk, 2016). We contend that this decision will, in practice, depend on a variety of factors including what the prior literature has shown, the

importance of confirming a specific relationship to the literature, the cost of pursuing multiple forms of research, and so on.

## Ethical and Privacy Considerations

The norms for ethical research and proper protocols in the context of popularized big data efforts are still developing, with some research studies generating considerable controversy over research ethics and participant consent (or lack thereof). Goel (2014) suggested that large-scale experiments introduce the potential for research conducted "on people who may never even know they are subjects of study, let alone explicitly consent" (p. 1). An illustrative example of this tension is the uproar caused when researchers scraped and then published data on nearly 70,000 users from the dating site OKCupid. Although the data were scraped from a pubic site, considerable controversy surrounded the publishing of identifiable data for research without consent (Leetaru, 2016). Even when researchers solicit participant consent, the interconnected and rich data of the big data era can bring about additional challenges. For example, obtaining Facebook data from participants (even those who have provided consent) inevitably results in collecting data on the friends of research participants who have not consented to being part of the study (similar concerns exist for Twitter, Fitbit, etc.). Should these data be off limits? If not, under what conditions is their use allowable? Open ethical questions also surround the use and solicitation of research participants on AMT and other crowd-sourcing platforms. In fact, some have termed AMT and the broader crowd-sourcing trend as "digital sweatshops," in which individuals are paid less than minimum wage to perform tasks without any rights or guarantee of payment if, for example, the requestor is not satisfied with workers' performance (Cushing, 2013).

The changing nature of data available for research also has implications for acceptable research practices. For instance, we have discussed how the rich set of variables available through big data efforts may exacerbate concerns about selective reporting of results. In particular, with an expansive and rich set of variables, concerns of post hoc hypotheses and spurious correlation may become more pronounced. However, the shift toward big data efforts could also diminish these concerns. Specifically, the abundance of study participants reduces concerns of underpowered studies and may actually help shift the focus away from statistical significance to practical significance and estimation of effect size magnitude. More so, claiming noneffects (i.e., showing support for the null hypothesis) may be more defensible with larger samples (e.g., in equivalence studies). In addition, predictive or exploratory efforts that leverage hundreds of variables and machine learning methods are not particularly conducive to isolating the specific impact of one variable on an outcome. In fact, many of these methods are considered "black box" in that they do not generate meaningful coefficient estimates for the individual variables included in them.

A related challenge is that research efforts using data from highly dynamic digital platforms where data is generated semi-continuously might pick up effects that are highly specific to a particular snapshot of data. While this does not necessarily mean that the results from these studies are incorrect or spurious, it may speak to the generalizability of observed phenomenon in big data studies. Some of the studies we have discussed address these types of concerns by replicating their findings across multiple samples collected at different times and sometimes also from different settings. For instance, Jones, Wojcik, Sweeting, and Silver (2016) used three small subsamples of Twitter users (~1,000 users) to analyze changes in negative emotion following three separate incidents of violent crime on college campuses. Similarly, Hibbeln et al. (2017) replicated their results across three different samples and experimental contexts. Also, Iliev et al. (2016) tested their hypotheses in two independent, time-stamped text corpora (Google Books and the New York Times).

Big data research in psychology may also elicit considerable privacy concerns. Many of the studies we have noted thus far involve the linking of public data widely available about millions if not hundreds of millions of individuals to highly intimate psychological dimensions of these individuals (personality, suicidal inclination, etc.). These kinds of predictions could have ambiguous (and sometimes detrimental) impacts on individuals if less well-intentioned entities were able to learn from this research and make the same type of prediction. For example, the Crystal Platform purports to be the world's largest "personality platform" and sells personality profiles (without a test) for individuals based on the analysis of "public data" about them. Some evidence points to commercial services that predict personality from social media data having a prominent impact on the 2016 U.S. election and the United Kingdom "Brexit" vote (Grassegger & Krogerus, 2017). Facebook, for example, recently prohibited a large car insurer in the United Kingdom from scouring customers' social media data to learn its customers' personality traits; the insurer hoped to charge different prices for premiums based on these insights (Rudgard, 2016). This raises difficult questions about what kinds of controls need to be in place to protect research subjects and to perform research ethically. These concerns are exacerbated by recent works highlighting consumers' limited ability to navigate complex privacy tradeoffs online (Adjerid, Peer, & Acquisti, 2017; Adjerid, Acquisti, Brandimarte, & Loewenstein, 2013). While an extended discussion of how to address these issues is outside the scope of this article, other works have discussed these issues at length (see e.g., Wienberg & Gordon, 2015; Boyd & Crawford, 2012).

## Discussion and Conclusions

Overall, we contend that psychology is well situated—in fact, more so than many other fields—to leverage the era of big data to advance research. In part, this is the case because a great deal of data collected in the big data context involves persons and often behaviors (e.g., purchasing decision, online activities, responding or reacting to emails, "likes" given to different companies or causes, status updates, written responses/comments) and therefore human processes. At the same time, big data research introduces a number of fundamental questions for and to the field.

First, varying technological and statistical constraints emerge as a consistent challenge for big data research efforts in psychology, as well as other fields. In particular, while developments at the intersection of computer science and statistics (e.g., machine learning) have opened up opportunities for research that addresses relevant questions to psychology using big data, utilizing and tailoring these methods for questions of interest to psychology is nontrivial. We contend, however, that psychology is well poised to tackle these challenges and in many ways already has experience with big data, from large-scale testing, to registries, to educational effectiveness evaluations. In fact, the fields of statistics and psychology also have a rich shared history that is still ongoing. As Stigler (1999) eloquently articulates, "statistics and psychology have long enjoyed an unusually close relationship . . . they are bound together" (p. 189). This includes the development of methods that have gained widespread appeal and popularity because of their generalizability to a range of problems. In fact, Gelman once noted in a talk, "I think we do know a longstanding principle in statistics, which has been rediscovered by some of the people here, which is that any idea you've had has already been done in psychometrics about 50 years ago."[2] Obviously there is some hyperbole in Gelman's statement, but the point is that the historical interdependence of psychology and statistics results in psychology researchers being highly versatile empiricists who are ready and capable to tackle the challenges leveled by big data and to reap the benefits it provides. In fact, we believe that researchers in psychology should pursue more development of methodological techniques for addressing psychology's unique challenges in this space (e.g., such as the recent special "Big Data in Psychology" issue of *Psychological Methods*, December 2016).

Although there is a path forward for psychological researchers using big data, challenges remain. At present, we fear that there may be a gap in the current methodological training in psychology and what is sometimes needed for big data research. Thus, undergraduate and graduate programs in psychology, particularly quantitatively focused ones, may consider expanding their methodological offerings, potentially by partnering with other disciplines such as computer science, information technology management, and business analytics to provide a more comprehensive suite of methodological tools that more readily address obtaining, organizing, and converting raw instrumented data into usable behavioral data. Similar to the review in the 1990s (Aiken et al., 1990) and again in the first decade of the 2000s (Aiken, West, & Millsap, 2008) both in this journal, it may be time for psychology as a field to review the methodological training with an eye on instrumented data, online generated data, advanced survey methods, et cetera. We see the need to supplement, not supplant what is currently taught, with offerings focused on data management techniques (e.g., web scraping, Hadoop for distributed computing) as well as approaches for analysis of data (e.g., machine learning, network analysis, and even artificial intelligence). Our experience is that it has long been common for doctoral students in business schools to take methodological courses from psychology departments. It may well be time for psychology students to take a similar approach and seek out courses that deal with instrumented data from business schools, particularly information technology management and business analytics programs, as well as computer sciences departments.

Relatedly, there is the question of whether big data research efforts will offer the same type of value as the methodological developments that have historically taken place in psychology? In particular, the multitude of ways in which big data can be parsed, raises novel measurement issues and highlights the need to revisit classical issues in a new context. This includes revisiting various types of validity, particularly discriminant and convergent validity. Many of the issues, for example, raised in Cronbach and Meehl's (1955) classic work on construct validity are still highly relevant, yet do not seem often considered in big data contexts (see also Messick, 1995, this journal). Measurement invariance (e.g., Millsap, 2011), for example, takes on new considerations do to different platforms in which one can interact in a digital environment (e.g., "platform invariance"). Of course, just become big data can be collected it is not necessarily the case that it is quality data or that the values map onto a meaningful psychological variable. As the president of the American Statistical Association recently noted, and a statement with which we agree, "in our field, in the age of Big Data, I am concerned we may have lost our grip on quality in favor of quantity" (Nussbaum, 2017). To make lasting contributions, sound research design and analysis considerations are important, regardless of the size of the data set. We would also like to point out that "small data" can have big value; there is no requirement that value is wedded to the size of a data set.

[2] Gelman, http://stat.columbia.edu/~martin/Workshop/statistics_neuro_data_931_speaker_04.mov.

Finally, important questions remain of whether big data insights will be of applied or theoretical value? These questions are even more relevant in the context of exploratory big data efforts, in which researchers seek to "understand what the data are saying" rather than develop post hoc hypotheses to explain why. Previously, we have argued that the myriad of opportunities that big data affords suggests that big data research will enjoy a symbiotic relationship with traditional psychological research (see section on "Theoretical and Research Value"). In this regard, journals and reviewers may want to consider being more open to data driven findings instead of a requirement by some that theory dictate what be examined. Nesselroade (2006) once discussed in the context of developmental psychology, although the ideas are much more general, how method and theory are like a dance. He noted that "theory has led the dance for a long time, perhaps long enough that it is time for a change [to allow method to lead]." At the same time, researchers should not assume that simply because their data is "big", that the insights they generate from it are interesting and publishable.

While it is clear big data research efforts come with unique challenges and risks that need careful consideration, there is much to gain from using big data and big data approaches to advance psychological research. Indeed, whereas these efforts have been thought to be beyond some researchers, we hope that we have shown that some applications are within reach and we hope researchers will consider if their work would benefit from wading into this territory and we hope inspire some researchers to do so. Meeting the challenges will not be easy but we believe can offer important steps for the advancement of psychology and related fields.

## References


Adjerid, I., Acquisti, A., Brandimarte, L., & Loewenstein, G. (2013). Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proceedings of the ninth symposium on usable privacy and security* (p. 9). New York, NY: ACM.

Adjerid, I., Peer, E., & Acquisti, A. (2017). Beyond the privacy paradox: Objective versus relative risk in privacy decision making. *Management and Information Systems Quarterly*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765097

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

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., Scarr, S., . . . Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45,* 721–734. http://dx.doi.org/10.1037/0003-066X.45.6.721

Andrzejewski, D., Zhu, X., & Craven, M. (2009, June). Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 25–32). New York, NY: ACM.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science, 337,* 337–341. http://dx.doi.org/10.1126/science.1215842

Bapna, R., Ramaprasad, J., Shmueli, G., & Umyarov, A. (2016). One-way mirrors in online dating: A randomized field experiment. *Management Science, 62,* 3100–3122.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66,* 153–158. http://dx.doi.org/10.1016/j.jesp.2016.02.003

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal, 38,* 218–226. http://dx.doi.org/10.1037/prj0000130

Boas, T. C., & Hidalgo, F. D. (2013). Fielding complex online surveys using rApache and Qualtrics. *The Political Methodologist, 20,* 21–26.

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. Cambridge, MA: MIT Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society, 15,* 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21,* 566–582. http://dx.doi.org/10.1037/met0000090

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. http://dx.doi.org/10.1177/1745691610393980

Cattell, R. B. (1946). Personality structure and measurement; the operational determination of trait unities. *British Journal of Psychology, 36,* 88–103. http://dx.doi.org/10.1111/j.2044-8295.1946.tb01110.x

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 67–128). Chicago, IL: Rand-McNally.

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

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *Management Information Systems Quarterly, 37,* 50–56.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: A general framework and some examples. *Computer, 37,* 50–56. http://dx.doi.org/10.1109/MC.2004.1297301

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65,* 145–153. http://dx.doi.org/10.1037/h0045186

Coombs, C. H. (1964). *A theory of data*. New York, NY: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. In *Flow and the foundations of positive psychology* (pp. 35–54). Rotterdam, the Netherlands: Springer. http://dx.doi.org/10.1007/978-94-017-9088-8_3

Cushing, E. (2013). *Amazon Mechanical Turk: The digital sweatshop. UTNE*. Retrieved from http://www.utne.com/science-and-technology/amazon-mechanical-turk-zm0z13jfzlin.aspx

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. Hoboken, NJ: Wiley. http://dx.doi.org/10.1002/0471448354

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., . . . Parmar, N. J. (2016). TACIT: An open-source text

analysis, crawling, and interpretation tool. *Behavior Research Methods, 49,* 538–547.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., . . . Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General, 145,* 366–375. http://dx.doi.org/10.1037/xge0000139

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47,* 1–12. http://dx.doi.org/10.3758/s13428-014-0458-y

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, 1,* 1542–1552. http://dx.doi.org/10.14778/1454159.1454226

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55,* 78–87. http://dx.doi.org/10.1145/2347736.2347755

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York, NY: Springer Science & Business Media. http://dx.doi.org/10.1007/978-0-387-49771-6

Garaizar, P., & Reips, U. D. (2014). Build your own social network laboratory with Social Lab: A tool for research in social media. *Behavior Research Methods, 46,* 430–438. http://dx.doi.org/10.3758/s13428-013-0385-3

Gentry, J., Gentry, M. J., RSQLite, S., & Artistic, R. L. (2016). *Package 'twitteR'*. Retrieved from https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

Goel, V. (2014). *As data overflows online, researchers grapple with ethics. The New York Times*. Retrieved from http://www.nytimes.com/2014/08/13/technology/the-boon-of-online-data-puts-social-science-in-a-quandary.html

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics, 20,* 830–851. http://dx.doi.org/10.1198/jcgs.2010.10007

Grassegger, H., & Krogerus, M. (2017). The data that turned the world upside down. *Motherboard*. Retrieved from https://motherboard.vice.com/en_us/article/how-our-likes-helped-trump-win

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer, 27,* 83–85. http://dx.doi.org/10.1007/BF02985802

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61–83. http://dx.doi.org/10.1017/S0140525X0999152X

Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S., & Weinmann, M. (2017). How is your user feeling? Inferring emotion through human-computer interaction devices. *Management Information Systems Quarterly, 41,* 1–21.

Hu, Y., Shmygelska, A., Tran, D., Eriksson, N., Tung, J. Y., & Hinds, D. A. (2016). GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nature Communications, 7,* 10448. http://dx.doi.org/10.1038/ncomms10448

Iliev, R., Hoover, J., Dehghani, M., & Axelrod, R. (2016). Linguistic positivity in historical texts reflects dynamic environmental and psychological factors. *Proceedings of the National Academy of Sciences of the United States of America, 113,* E7871–E7879. http://dx.doi.org/10.1073/pnas.1612058113

Iliev, R., & Smirnova, A. (2016). Revealing word order: Using serial position in binomials to predict properties of the speaker. *Journal of Psycholinguistic Research, 45,* 205–235. http://dx.doi.org/10.1007/s10936-014-9341-3

Jaffe, E. (2014). What big data means for psychological science. *Observer,* 27. Retrieved from https://www.psychologicalscience.org/observer/what-big-data-means-for-psychological-science

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York, NY: Springer.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532.

Jones, N. M., Wojcik, S. P., Sweeting, J., & Silver, R. C. (2016). Tweeting negative emotion: An investigation of Twitter data in the aftermath of violence on college campuses. *Psychological Methods, 21,* 526–541. http://dx.doi.org/10.1037/met0000099

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137–152. http://dx.doi.org/10.1037/a0028086

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759. http://dx.doi.org/10.1177/0013164496056005002

Koh, Y. (2014). Report: 44% of Twitter accounts have never sent a tweet. *Wall Street Journal,* 11.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist, 70,* 543–556. http://dx.doi.org/10.1037/a0039210

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 5802–5805. http://dx.doi.org/10.1073/pnas.1218772110

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 8788–8790. http://dx.doi.org/10.1073/pnas.1320040111

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods, 21,* 475–492. http://dx.doi.org/10.1037/met0000081

Leetaru, K. (2016). Are research ethics obsolete in the era of big data? *Forbes*. Retrieved from https://www.forbes.com/sites/kalevleetaru/2016/06/17/are-research-ethics-obsolete-in-the-era-of-big-data/#2353d8897aa3

Lohr, S. (2014). For big-data scientists, "janitor work" is key hurdle to insights. *The New York Times*. Retrieved from https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

Marder, J. (2015). The Internet's hidden science factory. *PBS News Hour*. Retrieved from http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory/

Mark, G., Iqbal, S. T., Czerwinski, M., Johns, P., & Sano, A. (2016, May). Neurotics can't focus: An in situ study of online multitasking in the workplace. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 1739–1744). New York, NY: ACM.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70,* 487–498. http://dx.doi.org/10.1037/a0039400

McFowland, E., III, Speakman, S., & Neill, D. (2013). Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research, 14,* 1533–1561.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). *Who is doing computational social science? Trends in big data research* (Whitepaper). London, England: SAGE Publishing. http://dx.doi.org/10.4135/wp160926

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science, 341,* 647–651. http://dx.doi.org/10.1126/science.1240466

Nesselroade, J. R. (2006). Quantitative modeling in adult development and aging: Reflections and projections. In C. S. Bergeman & S. M. Boker (Eds.), *Methodological issues in aging research: Notre Dame series on quantitative methods* (pp. 1–18). New York, NY: Erlbaum Associations.

Nielsen, F. (2016). Interview with Professor Calyampudi Radhakrishna Rao. *Amstatnews*. Retrieved from http://magazine.amstat.org/blog/2016/12/01/raointerview/?utm_source=andec16&utm_medium=email&utm_campaign=amstatnews

Nussbaum, B. (2017). President's Corner: Reflecting on Quality vs. Quantity. *Amstat News*. Boston, MA: American Statistical Association. Retrieved from http://magazine.amstat.org/blog/2017/06/01/reflecting-on-quality-vs-quantity/

Oswald, F. L., & Putka, D. J. (2015). Statistical methods for big data: A scenic tour. In *Big data at work. Data science revolution and organizational psychology* (pp. 1907–2800). New York, NY: Routledge.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23,* 184–188. http://dx.doi.org/10.1177/0963721414531598

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). New York, NY: Harcourt Brace.

Peer, E., Samat, S., Brandimarte, L., & Acquisti, A. (2015). *Beyond the Turk*. Retrieved from http://ssrn.com/abstract=2594183 or http://dx.doi.org/10.2139/ssrn.2594183

Purta, R., Mattingly, S., Song, L., Lizardo, O., Hachen, D., Poellabauer, C., & Striegel, A. (2016, September). Experiences measuring sleep and physical activity patterns across a large college cohort with Fitbits. In *Proceedings of the 2016 ACM international symposium on wearable computers* (pp. 28–35). New York, NY: ACM.

Ramo, D. E., & Prochaska, J. J. (2012). Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use. *Journal of Medical Internet Research, 14,* e28. http://dx.doi.org/10.2196/jmir.1878

Rhemtulla, M., & Little, T. (2012). Planned missing data designs for research in cognitive development. *Journal of Cognition and Development, 13,* 425–438. http://dx.doi.org/10.1080/15248372.2012.717340

Rivera, I. (2015). Package: 'RedditExtractoR.' Retrieved from https://cran.r-project.org/web/packages/RedditExtractoR/RedditExtractoR.pdf

Rudgard, O. (2016). Admiral to use Facebook profile to determine insurance premium. *The Telegraph*. Retrieved from http://www.telegraph.co.uk/insurance/car/insurer-trawls-your-facebook-profile-to-see-how-well-you-drive/

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to the new statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology, 66,* 47–54. http://dx.doi.org/10.1016/j.jesp.2015.09.013

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Mahwah, NJ: Scientific Research.

Shannon, G., Andrew, P., & Duggan, M. (2016). *Social media update 2016*. Washington, DC: Pew Research Center.

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods, 46,* 41–54. http://dx.doi.org/10.3758/s13428-013-0353-y

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24,* 1875–1888. http://dx.doi.org/10.1177/0956797613480366

Somanchi, S., Adhikari, S., Lin, A., Eneva, E., & Ghani, R. (2015, August). Early prediction of cardiac arrest (code blue) using electronic medical records. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2119–2126). New York, NY: ACM.

Stanton, J. M. (2013). Data mining: A practical introduction for organizational researchers. In *modern research methods for the study of behavior in organizations* (pp. 199–230). London, United Kingdom: Taylor and Francis. http://dx.doi.org/10.4324/9780203585146

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.

Troendle, J. F., & Mills, J. L. (2011). Correction for multiplicity in genetic association studies of triads: The permutational TDT. *Annals of Human Genetics, 75,* 284–291.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). *Online human-bot interactions: Detection, estimation, and characterization*. Retrieved from https://arxiv.org/abs/1703.03107

Wang, C., & Blei, D. M. (2011, August). Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 448–456). New York, NY: ACM.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., . . . Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 3–14). New York, NY: ACM.

Wang, R., Harari, G., Hao, P., Zhou, X., & Campbell, A. T. (2015, September). SmartGPA: How smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 295–306). New York, NY: ACM.

Wickhan, H. (2014). Tidy data. *Journal of Statistical Software, 59,* 1–23.

Wienberg, C., & Gordon, A. S. (2015, June). Insights on privacy and ethics from the web's most prolific storytellers. In *Proceedings of the ACM web science conference* (p. 22). New York, NY: ACM.

Xi, X., Keogh, E., Shelton, C., Wei, L., & Ratanamahatana, C. A. (2006). Fast time series classification using numerosity reduction. In *Proceedings of the 23rd international conference on Machine learning* (pp. 1033–1040). New York, NY: ACM.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 1036–1040. http://dx.doi.org/10.1073/pnas.1418680112

Received May 03, 2016
Revision received June 23, 2017
Accepted June 29, 2017 ■