Information and Organization 29 (2019) 41–56

Contents lists available at ScienceDirect

Information and Organization

journal homepage: www.elsevier.com/locate/infoandorg

Revisiting IS research practice in the era of big data

Steven L. Johnson*, Peter Gray, Suprateek Sarker

McIntire School of Commerce, University of Virginia, Charlottesville, VA, USA

ARTICLE INFO

Keywords:
Big data
Theory
Research methodology
Research challenges
Trace data

ABSTRACT

Through building and testing theory, the practice of research animates data for human sense-making about the world. The IS field began in an era when research data was scarce; in today's age of big data, it is now abundant. Yet, IS researchers often enact methodological assumptions developed in a time of data scarcity, and many remain uncertain how to systematically take advantage of new opportunities afforded by big data. How should we adapt our research norms, traditions, and practices to reflect newfound data abundance? How can we leverage the availability of big data to generate cumulative and generalizable knowledge claims that are robust to threats to validity? To date, IS academics have largely welcomed the arrival of big data as an overwhelmingly positive development. A common refrain in the discipline is: more data is great, IS researchers know all about data, and we are a well-positioned discipline to leverage big data in research and teaching. In our opinion, many benefits of big data will be realized only with a thoughtful understanding of the implications of big data availability and, increasingly, a deliberate shift in IS research practices. We advocate for a need to re-visit and extend traditional models that are commonly used to guide much of IS research. Based on our analysis, we propose a research approach that incorporates consideration of big data—and associated implications such as data abundance—into a classic approach to building and testing theory. We close our commentary by discussing the implications of this hybrid approach for the organization, execution, and evaluation of theory-informed research. Our recommendations on how to update one approach to IS research practice may have relevance to all theory-informed researchers who seek to leverage big data.

1. Introduction

The IS discipline was founded in an era when research data was difficult to gather and expensive to analyze, in sharp contrast with the situation today where the scale of data available is staggering. Based on current trends of data doubling every two years, there will be 44 times more data available in 2020 than there was in 2009 (Computer Sciences Corp, 2012). One major category of big data encompasses digital traces of human behaviors. Not only are more people online, but they are also undertaking a growing proportion of activities online. Everyday use of computers, mobile devices, and inexpensive sensors leaves copious trails of digital exhaust—data collected as a byproduct of human actions. What was once a trickle of data about individuals and organizations has become a torrent. And yet, even though they may no longer be warranted, core assumptions of data scarcity remain embedded in many research practices. In light of the tremendous increase in data availability, we believe that the time is ripe to revisit established approaches to theory-informed IS research.

IS academics have touted big data as an entirely positive development. A common refrain is: more data is great, IS researchers

* Corresponding author.
E-mail address: steven@virginia.edu (S.L. Johnson).

https://doi.org/10.1016/j.infoandorg.2019.01.001
Received 15 January 2018; Received in revised form 14 September 2018; Accepted 24 January 2019

know all about data, and we are a well-positioned academic discipline to prosper in this new era. Big data opens up new avenues for teaching, research, and collaboration even beyond our traditional home in business schools (Goes, 2014). And, certainly, with wide-ranging impacts on practice, education, and research, big data is a big opportunity for our field. It "is possibly the most significant 'tech' disruption in business and academic ecosystems since the meteoric rise of the Internet and the digital economy" (Agarwal & Dhar, 2014, p. 443). Indeed, IS departments are developing many new educational offerings to provide the skills required to support new organizational needs (Goes, 2014). But, there has been relatively little dialogue about how IS researchers might best take advantage of the availability of big data in our research efforts (c.f., McKenna, Myers, & Newman, 2017; Pentland, Pentland, & Calantone, 2017).

In this paper we explore the question: *how should IS research practices adapt to an age of big data?* High-quality IS research strives to better understand an IT-related phenomenon by analyzing "*what* is happening?", addressing "*who* cares?", and identifying "*why* is it happening?" Our concern is that without explicit consideration of how to conduct theory-informed research using big data, the IS field may gravitate towards big data empiricism that de-emphasizes the understanding of "*why*?" Our goal is to propose a research approach that incorporates consideration of big data—and associated implications such as data abundance—into one classic approach to building and testing theory. We hope our commentary will be relevant to all researchers who seek to leverage the availability of big data in their pursuit of theory-informed research.

2. Big data and IS research

Although there is no generally accepted definition of big data, one commonly cited characterization is the 3Vs model that emphasizes that data is now generated with greater volume, variety, and velocity than (at some unspecified point) in the past. Alas, the alliterative 3Vs mnemonic provides little direction for researchers who ponder how big data might fit into, or suggest changes to, existing research practices. In this paper, rather than trying to tightly demarcate what is—or is not—big data, we focus instead on key characteristics of big data that we believe may have the greatest impacts on theory-informed IS research. A foremost characteristic noted by Salganik (2018) is that big data is typically "created by companies and governments for purposes other than research" (p. 14). Both aspects of this characterization underscore key differences in big data vs. what is traditionally used in theory-informed IS research. The IS field has well-established best practices for how researchers can collect, evaluate, and analyze qualitative, quantitative, and experimental data created by researchers for the purpose of research. Less clear is how to balance the opportunities and short-comings of big data sources, but there are important new opportunities for researchers who can do so effectively.

One advantage of using data sources that were not collected by researchers is that data generated as digital exhaust from online activity (or from the use of digital devices) can, under certain conditions, more accurately reflect individual, group, and organizational characteristics and actions. Data collected in a passive, unobtrusive, and non-reactive manner may overcome social desirability bias and other distortions that occur when researchers directly interact with research subjects (Webb, Campbell, Schwartz, & Sechrest, 1966). Big data also lets us study new topics, develop and deploy new tools, and revisit old questions (Abbasi, Sarker, & Chiang, 2016; Agarwal & Dhar, 2014; George, Haas, & Pentland, 2014; Sharma, Mithas, & Kankanhalli, 2014). Data abundance is an "opportunity for social scientists that have heretofore been hamstrung by a lack of data" (Agarwal & Dhar, 2014, p. 445). Detailed data generated as a non-reactive byproduct of digital actions can also complement data collected by researchers through traditional methods of surveys and interviews. Indeed, fine-grained observations of individual and organizational actions are easier and cheaper to collect and analyze than ever before.

The difference in sources and purposes of big data is illuminated through the distinction between primary, secondary, and tertiary data offered by Kitchin (2014b, pp. 7–8):

- *Primary* data is generated by a study's authors within their research design for the specific purpose of answering the study's research questions.
- *Secondary* data was collected by someone else, for a different purpose, and is re-used by a study's authors.
- *Tertiary* data is new data derived from either primary or secondary data—for instance, counts, categories, and statistical results.

All three terms are interpretable only within a focal study, as primary data in one context can become secondary data in another. Further, their interplay can be complex, as different types of data are often combined. For instance, an innovation researcher might combine results of a survey of product development managers (primary data) with an index of organizational patents (tertiary data) that the researcher generated from a governmental archive (secondary data).

In traditional approaches to IS research, a researcher collects primary data and then transforms it into tertiary data via qualitative analysis or through statistical approaches such as factor analysis, principal components analysis, or simple aggregation. As Bowker and Star (1999) remind us, data is dependent upon the mental models and research approaches used to produce it; as such, when researchers collect and use their own primary data, it is framed within the thought system of the researchers involved, who controlled what was collected, interpreted, and used (Kitchin, 2014a). In contrast, big data researchers rarely collect primary data but, rather, typically start with computer-based extraction of secondary data which they then convert into tertiary data to meet their own needs. In a big data context, it is the tertiary data that is framed within the researchers' thought system, rather than the primary data.

Although there are advantages to incorporating big data sources into theory-informed research, there are also formidable challenges. Because big data is created for purposes other than research, data items rarely map cleanly to higher-order constructs that are the fundamental building block of IS theory. Even after researchers convert raw trace data into tertiary data suitable for analysis, big data still predominantly takes the form of what Schutz (1962) termed *first level constructs*, variables that represent common-sense thinking. This is different from primary data created specifically by researchers to reflect concepts that are higher-order and more abstract (Lee, 1991; Walsham, 1995). Thus, a key challenge for researchers reusing data generated for other purposes is to thoroughly understand the underlying paradigm, instrumentation, and context in which that data was collected, and how the data may have been transformed, aggregated, or decontextualized since collection (Kitchin, 2014b).

Given the major differences between big data and research-collected data, it is surprising how little discussion has arisen about how using big data should change the practice of theory-informed IS research. Some scholars have noted that the very nature of inquiry is likely to change, given that large data sets, advanced algorithms, and powerful computing capabilities can initiate and refine questions without human intervention (Agarwal & Dhar, 2014). Other commentators argue that the scientific method is likely to become obsolete, as with the "availability of huge amounts of data, along with the statistical tools to crunch these numbers … science can advance even without coherent models, unified theories, or really any mechanistic explanation at all" (Anderson, 2008). Perhaps "scientists no longer have to make educated guesses, construct hypotheses and models, test them in data-based experiments and examples. Instead, they can mine the complete set of data for patterns that reveal effects, producing scientific conclusions *without* further experimentation" (Prensky, 2009).

Conversely, researchers in closely related fields have warned of potential downsides of such thinking. Writing in *Administrative Science Quarterly*, Davis admonished: "The advent of big data, combined with our current system of scholarly career incentives, is likely to yield a high volume of novel papers with sophisticated econometrics and no obvious prospect of cumulative knowledge development" (G. F. Davis, 2015, p. 179). Another senior scholar lamented the possibility that big data might increase the risk that researchers will publish "unreproducible random noise" (Starbuck, 2016, p. 178). While predictions of the new model of science articulated by Anderson (2008) and Prensky (2009) above are provocative, and the alleged virtues of inductive machine-derived theories and "digital serendipity" are undoubtedly promising, they remain largely speculative (Kitchin, 2014b). Further, despite the increasing proclivity in IS and related fields to publish descriptive and predictive analysis of big data, we believe it remains important to offer a path for theory-informed scholars to leverage the interaction of theory and big data in ways that further the creation of generalizable cumulative knowledge.

3. Alternative approaches

To set the stage, it is useful to compare and contrast the strengths and weaknesses of two distinct yet related approaches to research: traditional IS research that follows the hypothetico-deductive approach, and an emerging approach appearing in organizational practices and in IS literature, which we term *big data empiricism*. In the first sub-section below, we start with an idealized description of the hypothetico-deductive approach, describe examples, and summarize critiques of this approach. In the subsequent sub-section, we describe key elements of big data empiricism, summarize recent published examples, and also summarize critiques of this approach.

3.1. Approach 1: classic hypothetico-deductive approach

Research based on hypothetico-deductive logic (left panel in Fig. 1) is intended to produce knowledge claims that are cumulative and generalizable (e.g., Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002). Over time, the specific research practices within this approach came to be optimized for a world of data scarcity. Research that follows this approach ideally begins by clearly identifying both a phenomenon of interest and corresponding research questions to guide scientific inquiry. This process is heavily informed by prior research on the phenomenon and by the researchers' adopted theoretical perspective. Next, evidence is gathered through a deliberate sequential process that starts with a theory-based model of constructs (and their proposed relationships), is followed by the identification of corresponding measures (for those constructs), primary data collection (of the measures), and, finally, analysis of the collected data.[1]

Early steps in this approach are designed to maximize the likelihood that laborious and time-consuming data collection efforts will produce data that is free from threats to validity that may limit or obviate its use as evidence in support of knowledge contributions. The subsequent analysis of that data addresses the research questions by providing empirical support for conceptual knowledge claims that offer new theory-based explanations for the phenomenon.

Thus, the hypothetico-deductive approach provides a powerful way of maximizing inference from scarce data by ensuring that each step is completed prior to moving on to the next, always with an end goal of generalizable knowledge. The rigorous specification of highly structured processes, decisions, and time-tested approaches helps address threats to validity (and thereby increase confidence that the data really will help answer the research question). For example, this approach privileges linearity in progress through its steps; indeed, in many cases, iteration back to prior steps risks undermining validity and reducing confidence in the results generated.

A first exemplar of the hypothetico-deductive approach is a recent AIS Best Information Systems Publication Award for 2015, "The longitudinal impact of enterprise system users' pre-adoption expectations and organizational support on post-adoption proficient usage" (Veiga, Keupp, Floyd, & Kellermanns, 2014) with elements corresponding to each component in Approach 1 of Fig. 1. Veiga et al. (2014) summarized insights from their ideation process through a careful review of the prior literature on system usage

[1] Normatively, these steps are strictly sequential. Although in practice iteration might occur, the normative ideal is perpetuated with sequential presentation dominating published IS research.

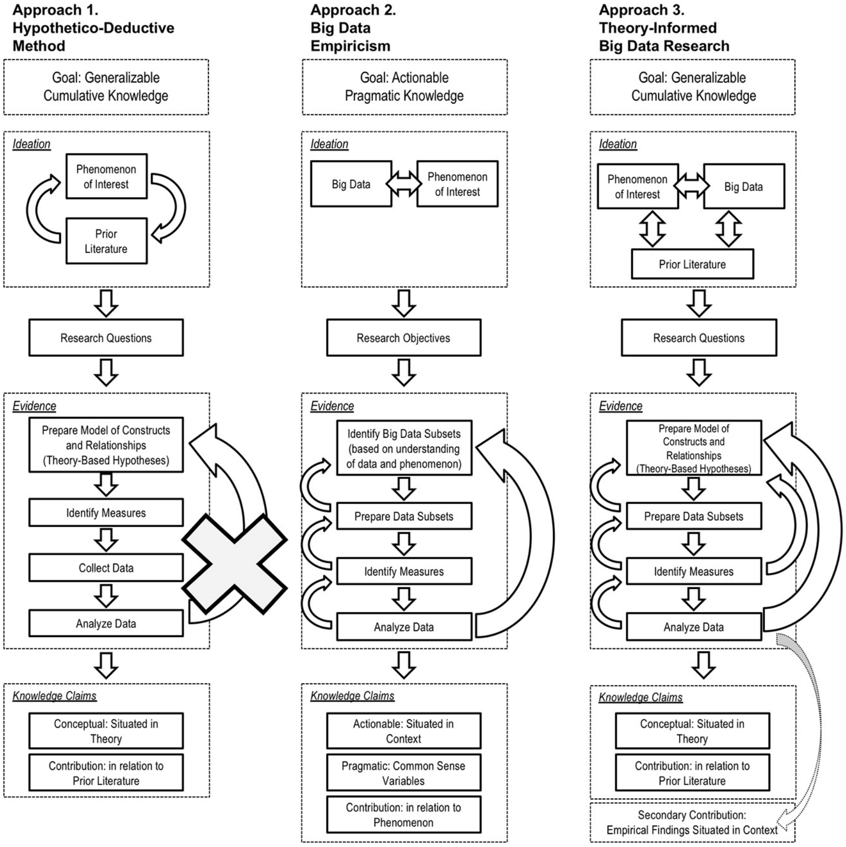

Fig. 1. Three idealized approaches to IS research.

that helped them identify key shortcomings. Their research question was to explain why "some adopters become highly proficient … while many other fail to leverage the benefits of even the most elementary applications?" (p. 692). They developed a concise theoretical model (p. 693) that captured both existing theory and their proposed extensions. Veiga et al. (2014) then laid out measures that optimally matched their theoretical constructs and described subsequent data collection (153 responses to two rounds of questionnaires) and analysis (a series of hierarchical regression models over 6 theoretical constructs and associated control variables).

Consistent with the hypothetico-deductive approach, Veiga et al. (2014) described their research as progressing in a linear fashion, with validity enhanced through careful efforts to ensure that the data analyzed could effectively address their research objective. As a result, readers (and presumably the Senior Scholars who selected it for this award) could have considerable confidence in this paper's knowledge claims—for instance, that adopters "who are internally motivated by stronger intentions to use [a system] are more apt to do so in ways that enhance their understanding and cumulative knowledge acquisition" (Veiga et al., 2014, p. 702), and that "when system adopters experience higher levels of organizational support, the indirect influence of pre-adoption expectations on proficient use is significantly higher." (Veiga et al., 2014, p. 703). We note that the research question as well as the contributions are stated in terms of higher level theoretical constructs, not in terms of every-day language of practitioners who were being studied.

We offer a second exemplar "Explaining variation in client extra costs between software projects offshored to India" (Dibbern, Winkler, & Heinzl, 2008) that uses a scientific case study research approach (Lee, 1989), also referred to as explanatory case study approach (Yin, 1994).[2] The authors described a theoretical puzzle: predictions of transaction cost economics (TCE) that have withstood empirical tests in many other contexts fail to explain why organizations that offshore projects to countries with huge wage differentials often fail to realize economic benefits predicted by TCE. To address this shortfall in TCE theory, Dibbern et al. (2008) developed a theoretical model that extends TCE by incorporating the knowledge-based view of the firm (KBV) where knowledge asymmetries between client and vendor are key to understanding project economic success, moderated by geographic, cultural, and language differences between client and vendor. Their model features seven propositions framed using theoretical constructs, which

[2] We would like to note that there are many genres of case research, including but not limited to the positivist case study, interpretive case study, critical-realist case study, and so on. While a large proportion of case research involves inductive/abductive approach, some well-known case studies utilize hypothetico-deductive logic to validate or invalidate theoretical propositions as in Dibbern et al. (2008).

Dibbern et al. (2008) test using data from six case studies they conducted. They provided detailed qualitative evidence from the case studies that map onto their theoretical constructs, and also constructed quantitative measures from this data (for instance, counting the number of quotes provided per finance employee in each company to arrive at a numerical score of extra costs that can be compared between companies; p. 349). Their analysis lead them to offer a range of knowledge claims about the usefulness of KBV as an extension to TCE, all of which was framed using theoretical constructs (such as absorptive capacity, asset specificity, and power distance) that have maximal meaning to academics but may be difficult for practitioners to fully appreciate, or act upon.

3.2. Critique of the hypothetico-deductive approach

Over the years, researchers have voiced a broad range of concerns with the hypothetico-deductive approach, which would require considerably more space than available here to summarize fully. However, we wish to highlight three relevant concerns. First, because research findings are articulated in theoretical terms, they can sometimes be difficult for managers to digest or act upon. Indeed, many of the rigor-relevance discussions and critical commentaries on academic IS and business school academic research, whether or not misplaced, have argued that such research may not effectively inform managerial practice. The second concern is that theory acts as a blinder and thus inhibits systematic or serendipitous learning about the phenomenon from the data—a concern that is magnified because operationalizations are often tightly tied to the theory being tested and because linearity in research approach discourages testing anything that was not hypothesized. A third concern is that when the study of a phenomenon reaches a point where constructs and associated research approaches are well established, researchers may be particularly prone to opportunistically pursuing “least publishable units” of research. That is, the hypothetico-deductive approach can facilitate making minor tweaks to existing models (for example, through the addition of a new construct as an independent variable, as a mediator, or as a moderator). When researchers further fail to incorporate each other's minor extensions, the result is an ever-expanding body of unintegrated and non-cumulative work (Gray & Cooper, 2010).

3.3. Approach 2: emerging big data empiricism

Next, we describe an idealized representation of big data empiricism, an emerging research approach directed towards producing actionable, pragmatic knowledge from large datasets (center panel in Fig. 1). Though this data-driven approach to research is not—to our knowledge—formally codified in research methodology texts (yet), through our reading of the literature and by interacting with data scientists and self-identified big data researchers, we have converged on what we intend to be a respectful, prototypical representation of its underlying features.[3] We hasten to add that big data research spans a variety of practices, including some that diverge significantly from what we describe. Our intent is not to conflate all big data research into this single approach; rather, it is to describe a realistic archetype of big data research that captures its essence as a primarily empirical and largely atheoretical endeavor.

Within this genre of work, research objectives often emerge by bringing together an understanding of available data and a related phenomenon of interest. That is, our investigations suggest that big data empiricism often begins when researchers obtain access to secondary data on a topic of relevance for practitioners. As a largely atheoretical approach, the collection of evidence involves iteration between selecting subsets of available secondary data, preparing them for analysis, identifying and constructing new tertiary measures, and analyzing these measures to determine their predictive power on an outcome of interest. Theoretical constructs are unlikely to drive the analysis, and conceptualization is therefore often not required, because the value of knowledge claims in big data empiricism derives from actionable predictions.

Big data, when used by quantitative researchers or by qualitative researchers using computational grounded theory approaches (e.g., Berente, Seidel, & Safadi, Forthcoming; Nelson, 2017), also provides an increased opportunity to analyze comprehensive data sets that may not be merely representative of a larger population but, rather, may contain data about all known members of that population. For example, studies of online behavior (e.g., open source software contributor, online community participation) can often leverage essentially complete data sets of participation within that focal setting. When a research project's goals are practical and empirical, predictive accuracy within a population may be more important than generalizing findings to other populations. In such a situation, it is productive for researchers to loop back to earlier phases of data selection, preparation, and measure identification at any time when they believe there is an opportunity to enhance the predictive power of their model. Some analytic software can automate much of this process and independently mine data for correlations, iteratively testing and refining models, and reporting only the most significant models that result. This sort of iteration through all four stages of evidence collection and analysis in Approach 2 of Fig. 1 enhances researchers' certainty in their knowledge claims because they are able to confidently state that repeated attempts to arrive at superior predictive models were unsuccessful. Once the best evidence that supports the research objective has been found, researchers report knowledge claims that are actionable and situated in context, often involving common sense variables whose meanings are derived locally. Importantly, the contributions are not made in relation to a generalized conceptual body of theoretical knowledge. They are in relation to the phenomenon, with scope defined by the data, helping practitioners with real-world challenges around the phenomenon of interest.

Next, we describe three studies consistent with a big data empiricism approach. To demonstrate the scope of this approach, we summarize examples from individual, product, and organizational levels. The paper “Internet's Dirty Secret: Assessing the Impact of Online Intermediaries on HIV Transmission” (Chan & Ghose, 2014), is an award-winning individual-level study that corresponds to

[3] Our description is also consistent with applications of big data analytics to IS research (e.g., Müller, Junglas, vom Brocke, & Debortoli, 2016).

Approach 2 in Fig. 1. Chan and Ghose (2014) described the increase in sexually transmitted diseases associated with the rise of online dating sites, and framed their research objective as estimating the size of the relationship between online classified dating site market entry and HIV progression (p. 956). Chan and Ghose (2014) constructed a tertiary dataset by incorporating secondary data from the Centers for Disease Control and Prevention and a 10-year history of market entry by Craigslist.org in 33 states. They constructed measures for key antecedents, outcomes and control variables, and assembled panel data together from 6 disparate datasets. Chan and Ghose (2014) reported a detailed set of analyses, and noted that their work was the first "empirical effort that quantifies the impact of site entry on increasing reported HIV cases" (p. 971). Their contributions—for example, that "the entry of Craigslist is related to a 15.9 percent increase in HIV cases" (p. 955)—and associated recommendations are targeted at practitioner audiences, specifically "policy makers and public health professionals" (p. 971).[4]

A paper by Geva, Oestreicher-Singer, Efron, and Shimshoni (2017), "Using forum and search data for sales prediction of high-involvement projects" also corresponds to Approach 2 in Fig. 1, but at the product level of analysis. Motivated by claims that social media data may lack accuracy in predicting offline economic outcomes, the authors use pre-existing data gathered by Google to augment traditional social media predictors, noting that "[a] common practice in the context of predictive research is to mitigate flaws in imperfect data by enriching the data with additional, meaningful, information." (p. 66). To predict sales of different automobile brands, they create a tertiary dataset that incorporates common sense variables such as past brand sales, gasoline prices, forum mentions, and volume of searches of each brand. Geva et al. (2017) use a modeling methodology that "is predictive, rather than explanatory," and explicitly disavow any theoretical contribution (p. 67). Their empirical work demonstrates that "predictive models based on inexpensive search trend data provide predictive accuracy that is at least comparable to that of more commonly used forum-data-based predictive models" (p. 68), and that accuracy increases greatly for value brands (vs. premium brands). Their results are clearly targeted towards managers in organizations that manufacture high-involvement products, and are framed around the relatively low cost and substantial benefit of using search trend data to improve sales forecasting (and thereby improve decision performance for a range of issues that hinge on sales forecasts accuracy, such as competitive analysis, inventory management, and supply chain optimization).

The paper "Towards a better measure of business proximity: Topic modeling for industry intelligence" by Shi, Lee, and Whinston (2016) provides an organizational-level example of Approach 2 in Fig. 1. Framed in the context of organizations' need to engage in competitive intelligence in high-tech industries where conventional (human) sources of industry knowledge may have limited effectiveness, Shi et al. (2016) propose a novel computational measure of business proximity that is grounded in the analysis of textual data from a pre-existing database of high-tech companies, people, and investors. Shi et al. (2016) conduct a variety of sophisticated statistical analyses, and validate their results against other practitioner classification systems in order to demonstrate superiority. The authors claim that their results "demonstrate the potential of extracting economically meaningful information from publicly available, unstructured data through large-scale computation" (p. 1036). Shi et al. (2016) hold that their approach has important implications for helping "entrepreneurs, venture capitalists, and analysts to navigate the constantly changing landscape of the networked business environment" (p. 1054). Indeed, they take their findings one step beyond mere claims, and embed their analysis in a tool to allows managers and analysts to "expedite the process of startup search and competition analysis as well as facilitates efficient new niche-market discovery" (p. 1049).

Together, these three examples show that big data empiricism has been successfully applied with a variety of data sources and levels of analysis. All three state their primary contributions in terms of practical, largely atheoretical, implications. The appearance of these papers in top management journals demonstrates that reviewers and journal editors value the findings generated with this approach.

3.4. *Critique of big data empiricism*

Despite the clear, powerful, and sometimes award-winning contributions that big data empiricism can make, it is not set up to produce cumulative theoretical knowledge. While this genre of research may lead to the development of situation-specific causal models, these models are not likely to generalize (Davis, 2015, p. 179). Moreover, the fact that big data empiricism focuses on first-order constructs limits its connection to the more abstract task of theory development. Further complicating the matter is the reality that big data is often created, owned, and processed by private organizations. As a result, it may be difficult if not impossible for academics to replicate and extend studies if the organizations controlling the data change how it is created, manipulated, and/or made available. Private ownership of data also accentuates the difficulty of pursuing critical research topics that may not be in the best interests of the organizations that own the data. While big data empiricism enhances the direct practical relevance of academics engaged in problem-solving for (and with) organizations, it inhibits our role as a neutral, dispassionate, and independent scholars providing the critical commentary on managerial practices that remains equally relevant to research and scholarship.

4. Comparison: role of threats to validity

Traditional theory-informed IS research is frequently concerned with establishing conceptual knowledge claims regarding causal

[4] Perhaps in response to the significant emphasis on theory in the *MIS Quarterly* editorial culture, Chan and Ghose (unlike many authors in other highly noted outlets that publish research of this nature) use theory for sensemaking of their results. Still, their study cannot be seen as contributing to abstract theoretical knowledge that is the hallmark of Approach 1.

inference—providing generalizable explanations for why and how things happen. A commonly used framework for evaluating such claims came from Cook, Campbell, and colleagues, whose highly influential books describe four categories of validity concerns: construct, statistical conclusion, internal, and external. Below, we describe how such threats to validity are addressed in the hypothetico-deductive approach and in big data empiricism.

4.1. Construct validity

Hypothesis testing has been the dominant approach to justifying knowledge claims in quantitative IS research. In the hypothetico-deductive approach, hypotheses are stated as relationships between theoretical constructs. For example, the first hypothesis of Viega et al., (2014), involves “pre-adoption intentions” and “proficient usage achieved post-adoption.” These higher-order, more abstract constructs are conceptualized as encompassing distinct, yet closely related, constructs (e.g., intentions: internal motivation, internal commitment) with each construct measured via multiple survey items drawn from the literature.

The relationship between a concept (identified as a construct) and how it is operationalized (via specific concrete measures) is critical to making knowledge claims. Construct validity concerns this relationship; it considers the strength of inferences made from indicators (operationalizations) to theoretical constructs (Shadish et al., 2002). Strong construct validity is achieved through theory-inspired instrumentation and thoughtful research design. Common threats to construct validity include inadequate explication of constructs, constructs that are confounded, biases in methodology that complicate inference, and a range of ways in which the research process itself may confound what is measured (Shadish et al., 2002).

In the traditional hypothetico-deductive approach, construct validity is supported through a series of steps that begin by carefully defining theoretical constructs and then developing or re-using instrumentation that clearly operationalizes the constructs. Efforts are made to ensure that constructs are not confounded with each other, and that operationalization will not introduce bias. Care is also taken to ensure that other aspects of the research method neither undermine the theoretical meaning of constructs nor introduce unwanted variance into their measurement. Along similar lines, Yin (1994) provides guidelines on ensuring different forms of validity in case study research.

In contrast, big data empiricism has virtually no interest in higher-order theoretical constructs because organizational data used in big data empiricism largely corresponds to variables that represent common-sense thinking. Thus, to the extent that a variable directly represents a common-sense construct, the traditional concerns of construct validity are largely irrelevant. For example, Chan and Ghose's (2014) variables include Number of HIV Cases, Craigslist Entry (to a community), and community demographic variables such as Median Household Income. The majority of their variables came from existing databases; that is, data that they did not collect. Their knowledge claims regarding the association of HIV transmission and introduction of Craigslist personal advertisements does not require conceptual constructs but, rather, is an association between common-sense variables or their aggregates that may serve as (or impersonate) surrogates of known constructs.

4.2. Statistical conclusion validity

Statistical conclusion validity refers to inferences about whether the observed covariation between variables is due to chance (Shadish et al., 2002). Researchers demonstrate statistical conclusion validity by following best practices in quantitatively assessing a range of issues that could undermine confidence in the outcomes of statistical tests. A study that has strong statistical conclusion validity presents evidence that reported covariation between variables is unlikely to be the result of the methodology employed—for instance, violation of assumptions behind statistical tests, unreliability in measures, or low statistical power (Shadish et al., 2002). Statistical conclusion validity is relevant to all quantitative research approaches; however, the hypothetico-deductive approach and big data empiricism diverge in how statistical conclusion validity manifests within the research process.

In the traditional hypothetico-deductive approach, statistical conclusion validity is strengthened by demonstrating that analyses have sufficient statistical power and by ensuring measured variables conform to assumptions of statistical tests. Measurement error is often quantified through the use of multi-item scales, and effect sizes are assessed to demonstrate practical significance. Treatments are carefully conceptualized and implemented in order to have only the theoretically intended effect.

A common threat to statistical conclusion validity is an over-reliance on *p*-values, which is particularly problematic if no adjustment is made for repeated tests. Also, when the collection of research data is expensive, fewer observations are gathered and limitations regarding the minimum ratio of measures to observations grow in salience. Thus, the gold standard for the hypothetico-deductive approach involves testing *only* relationships theorized in advance of data collection. Although this normative aspiration is not universally realized, iterative analysis of data weakens, rather than strengths, statistical conclusion validity.

In contrast, with very large data sets, big data empiricism faces different threats to statistical conclusion validity. With a large number of observations, it is possible to identify very small effects, and it is therefore even more important to assess practical significance rather than relying on arbitrary *p*-value thresholds. Likewise, big data typically comprises many measures; algorithmic mining of relationships among them, based on empirical rather than conceptual criteria, further opens the door to results that are statistically valid yet conceptually limited in terms of generating cumulative, generalizable knowledge. Moreover, researchers often make efforts to ensure that relationships discovered are meaningful and not spurious—for instance, through additional analysis that improve the robustness of claims by ruling out alternative causes. Thus, in big data empiricism the iterative data analysis often strengthens, rather than weakens, statistical conclusion validity.

4.3. Internal validity

Internal validity concerns inferences about whether observed covariation between constructs implies a causal relationship (Shadish et al., 2002). Studies that have strong internal validity have good defenses against plausible alternate explanations for why constructs are causally related. Typical areas of concern include characteristics of subjects, how they were sampled, how the study was designed, and how it was implemented (Shadish et al., 2002). Internal validity is threatened when there is good reason to believe that the results are an unintended artifact of the research process itself.

In the traditional hypothetico-deductive approach, whether in qualitative or quantitative studies, internal validity is strengthened by careful design of the research process to render unlikely any alternate explanations (beyond those proposed in a theory). Specifically, both in quantitative studies involving experimental setting as well as in longitudinal research design studies, whenever possible the research process is carefully designed to minimize confounding effects, subjects are selected or randomized to control for alternate explanations, and instrumentation is designed and tested to support the likelihood that causal connections are not a function of measurement itself. Quite often, however, researchers adopt cross-section designs or collect data in natural field experiments where major threats to internal validity may remain.

Big data empiricism addresses internal validity primarily through the use of statistical controls. Populations of subjects are largely given, based on what an organization has measured, and samples are selected on practical grounds. Big data is typically created with the sole purpose of meeting operational needs with no intention of being interpreted as higher-order concepts. When big datasets span periods of time, how variables are measured may also change in ways that introduce confounding explanations for any results found, as may subject maturation and attrition.

4.4. External validity

External validity refers to inferences about whether causal relationships can be generalized to different populations, settings, and times (Shadish et al., 2002). Studies that have strong external validity provide evidence that the sample and research setting are representative of the broader context to which a theory is intended to generalize. Common threats to external validity include the use of samples, measurement approaches, and settings that may be different in important ways from what is theorized (Shadish et al., 2002).

The traditional hypothetico-deductive approach often seeks to enhance external validity through deliberate sampling strategies and qualitative descriptions of samples to support the claim that they do not differ from the broader population to which a theory is intended to generalize. In some cases, key theoretically-important sample characteristics are measured in order to provide empirical evidence of their representativeness. In other cases, weaker approaches are used: for instance, showing that subjects' distribution of general characteristics such as age, gender, or education level are not radically different from what is contemplated in a theory, and describing research settings as a way of imply generalizability of results to other similar settings.

The issue of external validity often takes on a much narrower focus in big data empiricism. Organizations that mine their own data explicitly seek results that matter *within their own organizational context*, and therefore it does not matter whether results would also apply to a different population or in a different setting. A key objective of big data empiricism is to derive results that will guide future action, and as a result temporal generalization to later points in time is important. However, because big data empiricism generally seeks to produce practical knowledge that can be acted on to improve some metric of interest, few efforts are made to even understand external validity in the broader sense, let alone enhance it.

5. Charting a path forward

Many of the threats to validity that are highly salient to the hypothetico-deductive approach seem to be less of a concern to big data empiricism. We believe that this is due to major differences in goals associated with each approach. The traditional hypothetico-deductive research approach was developed over several decades when research data was scarce and took considerable time and effort to collect, and where a single critical threat to validity discovered during the publication process could derail years of work. Big data empiricism, ascendant in a period when data is abundant and readily available to researchers, has more pragmatic aims.

We return now to our primary concern and ask: *how can IS researchers incorporate the abundance of big data into their normative process of theory-informed research?* It is clear that Approaches 1 and 2 discussed above differ along many dimensions important to the practice of research. As summarized in Table 1, the hypothetico-deductive approach and big data empiricism differ in what makes for interesting research questions, in standards for rigorous evidence and inference, in the role of theory in the research process, and in value judgments regarding knowledge claims.

In our opinion, neither is inherently superior—both can generate valuable insights and advance understanding of phenomena of interest. Nonetheless, we also believe that just as the traditional hypothetico-deductive approach is poorly suited to embrace big data, big data empiricism is limited in its ability to integrate with the accumulated knowledge base of social science theory. Thus, we advocate for the formulation of hybrid approaches that adapt the goals and standards of the hypothetico-deductive approach while also leveraging the benefits of big data empiricism. Our goal is to capture the best of both approaches in the hybrid form described below, which we refer to as theory-informed big data research.

Table 1
Contrasting the two approaches.

	Approach 1: Hypothetico-deductive approach	Approach 2: Big data empiricism
Where do interesting research questions come from?	● Conceptual gaps, contradictions and frontiers based on phenomenon and prior literature Research questions limited by data availability ● Research questions limited by data availability	● Pragmatic research aims based on phenomenon and practice ● Research aims enabled by data availability (though sometimes prone to "street light effect")
What kind of data?	● Primary data ● Deliberately generated samples ● Data seen as evidence consistent with or inconsistent with a theoretical standpoint.	● Secondary data ● Pre-existing data with *n* approaching all (exhaustive "exhaust") ● Data seen as true and comprehensive representation of the phenomenon of interest
What are standards of inference?	● Sample sizes determined by statistical power ● Standards stable, though recently being questioned	● When *n* is very large, sufficient statistical power is assumed ● Standards unclear and gradually emerging
What is the role of theory in a research effort?	● Theory-phenomenon interaction at the heart of research process ● When data is scarce, theory enhances validity of expensively created data	● When data is abundant, theory is optional in research process ● Piecemeal theoretical reasoning can help identify plausible connections between variables in order to rule out spurious correlations
What is the primary contribution?	● Conceptual claims situated in theory and made in relationship to prior literature	● Actionable claims situated in context and made in relation to a better understanding of a phenomenon

5.1. Approach 3: Theory-informed big data research

As the hypothetico-deductive approach and big data empiricism support distinct practices and are optimized for different outcomes, can they be merged to leverage the opportunities provided by data abundance without sacrificing the goals of theory building and cumulative knowledge generation? We believe that a thoughtfully constructed hybrid process for building theory with big data (right panel in Fig. 1) can best leverage each of their strengths. As summarized in Table 2, theory-informed big data research adopts elements of both approaches. Similar to the hypothetico-deductive approach, the goal of theory-informed big data research is to generate conceptual knowledge claims that result in cumulative and generalizable knowledge. Similar to big data empiricism, the research process begins with an investigation of available data, rather than the presumption that new, researcher-generated data is necessary. As detailed below, this hybrid approach maintains consideration of traditional validity concerns while also supporting theory-informed iteration in data analysis. Next, we offer a brief overview of theory-informed big data research and an example to show how it might be conducted before discussing detailed implications associated with this hybrid approach.

In considering how to incorporate the availability of big data into project ideation, we note that data volume is not the primary challenge; after all, some of the historically largest data sets, such as U.S. Census Bureau data, have been successfully analyzed by social scientists. Instead, a primary challenge for using big data in research is that "much of what is generated has no specific question in mind or is a by-product of another activity" (Kitchin, 2014a, p. 2). Further, this digital exhaust is often not even intended for direct human interpretation but, rather, exists to facilitate communication among computer programs. Thus, whereas census data is carefully created specifically for use by researchers, big data is certainly not.

To avoid the streetlight effect of initiating research just because a big dataset exists (Rai, 2016), we argue that it remains important to prepare research questions by triangulating across available data, associated phenomenon, and relevant prior literature offering candidate theoretical perspectives. To generate theory-informed research questions in this way, this phase critically involves a careful examination of available data, considering its representativeness of the phenomenon of interest and its potential to contribute to theory. This initial diligence is essential to fulfill the ultimate objective of building cumulative knowledge.

The next phase in this hybrid approach is to build a detailed conceptual model, informed by the chosen theoretical perspective

Table 2
Summary of theory-informed big data research.

Where do interesting research questions come from?	● Tractable theory-informed research questions are generated through triangulation of understanding the phenomenon, available data, and prior literature ● Most interesting research is based on novel data sources that can address conceptual gaps, contradictions and frontiers
What kind of data?	● Tertiary data (summary of primary or secondary data)
What are standards of inference?	● When *n* is very large, sufficient statistical power is assumed ● Theory-informed data subsets can enhance internal and external validity
What is the role of theory in a research effort?	● Theory appears in beginning, middle, and end: o Theory helps prioritize what big data is worth attention o Theory guides process of turning big data into manageable subsets o Theory informs what is (and is not) a contribution
What is the primary contribution?	● Conceptual claims that are situated in theory and made in relationship to prior literature

Table 3
Example of applying three approaches in similar research projects.

Approach	Hypothetico-deductive approach	Big data empiricism	Theory-informed big data research
Ideation	Phenomenon: Enterprise Social Network (ESN) usage and sales team performance Research Question based on gaps and potential to enhance theory: do team perceptions of ESN predict impact of ESN usage on manager ratings of team performance?	Phenomenon: sales teams usage of ESN Big data: detailed ESN usage logs over a multi-year period, coupled with human resource and sales performance data Research Aim: predict sales team performance	Phenomenon: ESN usage and sales team performance Big data: detailed ESN, human resource, and sales performance data Research Question based on the relevance of available data to several candidate theories: what ESN usage patterns are most impactful on team performance?
Evidence	Design and test instrument based on past literature. Conduct longitudinal survey of sales team members. Collect team perception data using constructs established in the literature and theoretically relevant control data at beginning of period; monthly questionnaires for usage and perceptions thereafter. Empirically test hypotheses only.	Import all available data into analysis environment. Run machine learning classification of team ESN usage, team and employee characteristics; and performance outcomes. Iteratively test many combinations of predictors until algorithm can find no better model.	Prepare model of constructs and hypothesized relationships based on published theory. Examine metadata to identify variables that are conceptually similar to those proposed in theory. Discard variables that correspond only weakly, and revisit model if metadata suggests operationalization may be problematic. Converge on measures: use of closed channels (within-team-only), use of open channels (cross-team specializations), and performance outcomes.
Possible knowledge claims	Initial team perceptions of ESN (expectation-confirmation) predicts impact of usage on team performance.	Teams with highest performance relative to manager expectations are those where the junior-most members of a team are the most extensive users of the ESN.	Boundary spanning (intensive use of open, cross-team channels) enhances team performance when there is also strong team cohesion (intensive use of closed, within-team channels). By-products from the iterative analysis that may have practical implications: open channel use can be detrimental to team performance if there is no corresponding closed channel communication.

that aligns with available data. This requires a thorough understanding of the structure and history of the secondary data, but it is a separate and distinct step from analyzing the data itself—more akin to analyzing metadata than to performing statistical analysis. Iterating between conceptual models that could be tested and related measures that could operationalize them leads to a robust model with potential for theory building. In the following step, analyzing data to test the model, numerous big data analysis techniques may be fruitfully leveraged. Although practitioner-relevant knowledge may result from empirical analysis (and, thus, provide a secondary research contribution), overall this approach accepts reduced precision in localized prediction as a necessary tradeoff for creating interpretable models of explanation. Finally, echoing the hypothetico-deductive approach, evidence is presented to support conceptual knowledge claims with the goal of further explaining the phenomenon and addressing the identified research questions embedded within a theoretical tradition. A similar though not identical approach has been proposed for computational grounded theorists consisting of three phases: “pattern detection” using “unsupervised learning”, “pattern refinement” through “interpretive engagement” with “sociologically meaningful concepts”, and “pattern confirmation” that involves deductive assessment (Nelson, 2017). In summary, this hybrid approach shares similar goals and outcomes with the traditional hypothetico-deductive approach, while also incorporating some of the key advantages of big data empiricism.

5.2. An example demonstrating the value of the proposed hybrid approach

To illustrate the distinctive nature of these three approaches, consider different research projects that seek to understand when sales team performance is enhanced by enterprise social networks (ESN). We note that because of the wider applicability of big data and hypothetico-deductive approach to quantitative research, our examples are deliberately chosen to be quantitative in nature.

In Table 3, we summarize three hypothetical research projects that map onto the approaches summarized in Fig. 1. In a study adopting the traditional hypothetico-deductive approach, project ideation is driven by the phenomenon (ESN usage to enhance sales team performance) and existing theory. A big data empirical approach is driven by available data and the phenomenon that data represents—in this example, detailed ESN, human resource, and sales team performance data. In a hybrid approach, available data and associated phenomenon are considered in relation to a theoretical-informed research question drawn from social network theory.

As illustrated in these three hypothetical research projects, each may have similar aims (e.g., understanding how IT enhances team performance), but differ in key activities and in the nature of ensuing knowledge claims. For instance, specific activities in evidence generation differ, as does the amount and type of iteration possible. Approach 1 avoids iterating at all costs, with a structured and linear approach through activities, while Approach 2 iterates intensely during the evidence-generation stage to identify the model that has the highest predictive power. In contrast, Approach 3 iterates to identify the best measures and data subsets that could test the model in question, and might need to revise the research model if the data turn out to be poor operationalizations of intended theoretical constructs. Knowledge claims also differ significantly; at two ends of the spectrum, the hypothetico-deductive approach generates knowledge claims with a strong conceptual basis while big data empiricism generates highly pragmatic, actionable knowledge claims. In contrast, the hybrid theory-informed big data approach generates a knowledge claim that is conceptually informed (a theoretical contribution building on past research) with practical implications (highly actionable findings that are a function of naturally occurring behaviors). The in the sections below, we describe this hybrid approach in detail, along with associated unique benefits.

5.3. Research project ideation

Moving from data scarcity to data abundance, neither the hypothetico-deductive approach nor a purely inductive form of empiricism is well-suited for identifying IS research projects worth pursuing. The hypothetico-deductive approach engages with data far too late in the research process to derive inspiration from the abundance of large datasets, while big data empiricism lacks the fundamentally creative engagement between theory and meta-data whereby researchers imagine new theoretical challenges and construct theory-guided research questions and hypotheses that can be answered by leveraging existing data.

Our proposed hybrid approach takes project ideation as a synergistic interaction between (a) phenomena that are the focus of researchers' interests, (b) research questions generated through a deep understanding of published theory combined with researchers' own novel creative insights, and (c) the availability of big data sources. The more researchers can learn about characteristics of potential big data sources, the more they will enhance their essential knowledge of the phenomena, the better they can identify appropriate theory, and the more likely they are to develop impactful research questions. Reciprocal interactions across these three components produce novel ideas for advancing theory through a strong alignment between what is known, what is imagined, and what is measured.

This approach contrasts with big data research methods described by Agarwal and Dhar, in which “the computer becomes an active question asking machine” that may even “initiat[e] interesting questions and refin[e] them without active human intervention” (2014, p. 444). Though algorithmic mining of datasets for correlations could generate important new knowledge about empirical relationships, it cannot generate or develop rich social science theories that have broad sets of implications. At best, it might reveal novel patterns of correlation—albeit, at enormous risk of Type 1 error if the computer has been hard at work testing every possible correlation—that researchers might then attempt to explain in subsequent hypothetico-deductive studies (e.g., Shmueli, 2010). While this algorithm-driven approach may produce surprising results that may even be counter-intuitive—exactly the interesting sort of thing that is valued by reviewers (Davis, 1971)—it is unlikely to produce a cumulative body of theory in the social sciences (Davis, 2015).

Despite the possibility that big data empiricism could aid project ideation through the identification of anomalous empirical results that warrant further exploration, even in this situation, it would only be through the interaction of data, theory, and an understanding of prior literature that researchers could identify what is truly anomalous or surprising. Project ideation is a fundamentally creative task, where researchers with deep conceptual understanding identify promising avenues for productively extending, revising, or replacing existing knowledge. Regardless of whether the goal is to generate cumulative or ground-breaking knowledge, insights into what questions are worth pursuing are beyond the capacity of current artificial intelligence, and are more likely to occur through an integration of big data abundance, researcher ingenuity, and immersion in a phenomenon.

5.4. Accumulating evidence

With a research question tied to one or more big datasets firmly established in the ideation phase, researchers can proceed to an iterative process of articulating a theoretical model and specifying the measures necessary to test it. This process is iterative because some variations of theoretical models may not be testable using the identified big dataset, but this can only be determined through efforts to operationalize those models using that specific data. The tight coherence between theory and measures that is necessary in the hypothetico-deductive approach (Bacharach, 1989) here is accomplished in a different way; rather than designing instrumentation that is consistent with a set of theoretical constructs, researchers must instead explore what is possible with big data and discover the operationalizations that best fit their theoretical model. Importantly, this exploration process does not involve testing associations between variables—the goal is not to improve prediction, as is the case in big data empiricism. Instead, the goal is to establish those measures that are the best operationalization of the desired theoretical constructs, and that provide the best defense against plausible threats to validity.

In the proposed hybrid approach, it is only after a theoretical model and operationalization are established that researchers proceed to the next step: empirically assessing the explanatory power of a theoretical model using statistical techniques that are sensitive to the unique issues in big datasets (e.g., Kaplan, Chambers, & Glasgow, 2014). Though not a radical departure from traditional research, our re-sequencing of activities is key to taking advantage of big data.

5.5. Asserting knowledge claims

The hypothetico-deductive approach maximizes the likelihood of producing generalizable knowledge claims, ones that are essentially about theoretical concepts and are cumulative across studies. In contrast, big data empiricism is intended to produce highly pragmatic, locally actionable knowledge claims that are not particularly transferable across contexts. There is a clear tension here in the nature of the contribution to knowledge that is anticipated by each approach.

We believe that the nature of knowledge claims offered by the hybrid approach stands a better chance of producing the kind of impact envisioned by Lewin (1945) when he famously argued that nothing is so practical as a good theory. Moreover, the hybrid approach is not blind to un-hypothesized insights embedded in the data to the extent the traditional hypothetico-deductive approach may be. Indeed, the hybrid approach opens up new possibilities for inductive learning through the generation of additional practitioner knowledge as a byproduct of the research process—something that could also help make our research more accessible and valuable to external stakeholders. Anchored in theory and using datasets drawn from operational contexts that would be impossible for researchers to construct on their own, theory-informed big data research knits together the rigor of our classic research tradition and the relevance of phenomena and data that are important for key practitioner constituencies.

6. Implications for IS research practice

Next, we describe implications of adopting a theory-informed big data research approach.

6.1. Incorporating big data into research project ideation

In theory-informed big data research, the project ideation phase requires a careful balance of three sources of inspiration: a phenomenon, a body of published research that includes but is not limited to a theoretical perspective, and the availability of big data. Each is individually important, and the relationship among the three equally so.

6.1.1. Don't let data availability dictate the phenomena we study or the questions we ask

Because big data is typically captured by non-researchers, it is the result of a priori decisions made outside a research context—e.g., managerial priorities and perspectives, organizational strategies, and choices reflecting limitations (technological and otherwise) on what can be collected. Big data is generated primarily in support of commercial activities, in order to address objectives that are quite different from academic concerns. Such byproducts of commercial activity may, therefore, be ill suited for theory development in a range of ways—for instance, the likelihood that data will not match operationalizations established in past research, and that datasets will be missing key predictors established as theoretically important. Research driven by the mere availability of big data thus risks producing idiosyncratic, non-cumulative outcomes. More broadly, it privileges managerial decisions over informed researcher judgments; if we only study what managers think should be measured, our research will lose its ability for critical insight, its potential to reveal and emancipate stakeholders from faulty assumptions or even false consciousness, and its contribution to cumulative theory development.

These risks do not doom big data as irrelevant to the academic enterprise directed towards developing valuable theories; rather, caution is warranted. When data was scarce, the rare availability of large organizational datasets was worthy of researchers' consideration; now, researchers' attention has become a scarce and valuable resource. Thus, not all data merits study by IS researchers. Furthermore, the mere act of prospecting for big datasets can broaden our understanding of interesting phenomena and, thereby, further enhance theory development. To maximize the contribution of our research, we should be choosy about what data is worth studying.

6.1.2. Immerse deeply in the phenomena to understand the context of big data

The automated processes that organizations use to create, collect, and organize large amounts of data are based on countless fine-grained operational trade-offs. These organizational decisions are made by different individuals, at different points in time, with different levels of intentionality and documentation. As such, there is no single informant capable of providing a researcher with every critical contextual detail. Big data sources have a history, including changes in subjects, technology, structures and algorithms by which data were calculated, imputed, or measured, and the processes by which data was captured, coded, cleaned, merged, and manipulated. Moreover, decisions about data handling and manipulation will have changed over the lifespan of a dataset such that the meaning of a variable may not be consistent over time. As a result, big data that is made available to academic researchers is often divorced from the contextual knowledge required to understand its relationship with real world phenomena.

When organizations analyze their own data, these issues are less problematic. After all, an organization that generates data regarding its own products, customers, and processes possesses the institutional knowledge required to produce actionable analyses. In contrast, academic researchers are not embedded in the data-generating organization and have different objectives for analyzing it. Thus, another important consideration for IS researchers when evaluating potential big data sources is the extent to which necessary contextual knowledge is available—not just of the associated phenomenon writ large, but, for a specific data source.

By deeply immersing themselves within a phenomenon, researchers are better situated to establish the authenticity and representativeness of available data. For example, in discussing qualitative analysis of social media, McKenna et al. note “with user generated data … there is less control and less knowledge about the origin of the data, meaning there is potentially much more noise in the data (irrelevant data) which needs filtering” (McKenna et al., 2017, p. 90). This is another way in which researchers need be choosy—we must be willing to walk away from potential data sources that lack supporting contextual knowledge regarding their provenance and applicability.

6.2. Handling complex big data created by others

For a traditional IS researcher, analyzing large complex datasets created by others creates a new set of challenges. We discuss implications of this aspect of big data below.

6.2.1. Take action to counter plausible threats to validity in big data

A defining characteristic of big data is that it was created for purposes other than research. Compared to data generated by researchers following an intentional pre-planned research design, big data are especially at risk of a number of threats to validity. Subjects in big datasets are rarely random subsets of the population to which theory is intended to generalize, and may be significantly different from researchers' populations of interest in many ways, including: differences in demography, psychology, attitudes, and beliefs; situational differences in terms of the organizational treatments to which they have been exposed; differences in history of interaction with an organization; differences in economic, social, and technological environments. There is no easy solution here, but because these differences may be difficult to assess, more effort must be expended to fully understand exactly what population a sample is likely to generalize to, either through organizational informants or via additional data collection from subsamples to quantitatively assess departures from populations on key dimensions (and perhaps identify subsamples that are more representative).

Big data has large volume in part because it often spans significant periods of time. The more time a dataset covers, the more metadata characteristics may change, including the methods by which the data was generated, the algorithms by which variables were calculated, and the IT artifact experienced by subjects. Identifying historical conditions and methods and then controlling or correcting for them in data that extends over many years may be particularly challenging (if not impossible) to do, particularly if organizations do not or cannot explain how variables have evolved over time (e.g., Google PageRank). To minimize such confounding influences, slicing big data into subsets that span shorter time periods and replicating analyses over several subsets may boost confidence that such confounding effects are unlikely.

Secondary data is also unlikely to conform to known and validated instruments and measures used by researchers, and construct validity often cannot be assessed using traditional techniques. The thought systems and interests implicitly embedded in big data instrumentation derive from operational, rather than conceptual, needs. Data that is collected by non-researchers is therefore likely to produce variables that are difficult to demonstrate as having a good fit with theoretical constructs. In such cases, other approaches for demonstrating construct validity may have value—for instance, collecting additional data using theory-informed measures from a sample of subjects in big datasets that can be used to validate big data empirical measures. This is in line with the increased appreciation in the IS discipline of the virtues of multi-method studies (e.g., Venkatesh, Brown, & Sullivan, 2016) and with recommendations for qualitative researchers studying big data to gather multiple sources of data and use mixed methods to triangulate their data (McKenna et al., 2017).

6.2.2. Engage big data using multi-disciplinary teams

Since the beginning of the IS field, there has been a steady upward trajectory in the number and complexity of skills required to execute and publish A-tier research (and a corresponding increase in the average number of authors per published paper). Theory-informed big data research requires broad skillsets such as data handling, manipulation and complex statistical analysis. At the time this manuscript was written, anecdotal evidence from the IS faculty job market suggests that only a small minority of IS faculty have high levels of competency in the computational, algorithmic, data-intensive skills required to perform sophisticated big data empirical analyses. Instead of seeing the acquisition of data science skills as a significant hurdle, IS researchers who are interested in theory-informed big data research but who lack such skills may instead consider forming larger research teams with additional expertise. Just as when the appearance of novel statistical techniques in our field required deep expertise in that technique, the abundance of big data suggests that some research teams may discover new opportunities to develop powerful new theory by enlisting data scientists as coauthors.

Likewise, researchers with the big data processing and analysis skills can benefit from including theory-informed researchers from other traditions. For example, there are at least three ways in which qualitative researchers can contribute to research using big data. First, qualitative researchers are often quite experienced in moving from data to constructs—a skill much needed in our proposed hybrid approach, too. Second, a large proportion of big data is unstructured and is amenable to traditional qualitative approaches for analysis (c.f., McKenna et al., 2017), providing the opportunity for triangulation or other synergistic efforts. Third, qualitative researchers are often experienced in deep immersion in a phenomenon and can apply those skills to better understanding the providence and representativeness of data. Indeed, qualitative methods for directly interacting with individuals (such as surveys and interviews) may be essential for understanding the relationship of collected big data back to the phenomenon of interest. The presence of qualitative researchers on a multi-disciplinary theory-informed big data research project is thus likely to enhance its success in a variety of ways.

6.2.3. Document methods before performing analyses

There is a growing awareness that researchers' choices about how to analyze data may unintentionally lead to capitalizing on chance (Type 1 error) when those choices are contingent on the data itself (Gelman & Loken, 2014). No matter the size of the dataset, researchers increase the odds of capitalizing on chance when making decisions about research models (e.g., control variables, cutoffs, moderators, variations in causal models) after interacting with the data. While this risk exists with many quantitative research approaches, it is magnified with big data.

Large datasets exist prior to researchers having fleshed out a plan for analysis, and often include many more indicators than when data was researcher-generated; as a result, there are far more opportunities to make seemingly-reasonable decisions after interacting with data that unintentionally capitalize on chance. The large number of indicators in big datasets also makes it possible to test many different models predicting an outcome variable, perhaps even using completely different sets of antecedents drawn from the same dataset; researchers who do so without adjusting for multiple comparisons are capitalizing on chance, even if each model is published separately (Gelman & Loken, 2014). Researchers who plan and document their decisions regarding model construction and testing prior to testing their hypotheses can reduce the risk of inadvertent Type 1 error, and also make their study easier to replicate (by themselves or by others). The use of statistical tests that adjust for multiple comparisons also seems prudent. Where there is no strong precedent for specific analysis decisions, post-hoc testing for sensitivity of results to alternative decisions enhances robustness of findings.

6.3. Conceptualizing and contextualizing big data

Compared to big data empiricism, the objective of our hybrid approach is to raise the level of conceptualization such that the interpretation of big data can be situated within and contribute to IS theory. This requires close attention to the issues described below.

6.3.1. Don't confuse actions with attitudes

A rich tradition in the IS literature investigates attitudes and beliefs as predictors of actions and ultimately outcomes, where in many cases some or all data was self-reported. In this historical research context, data on actual actions and outcomes was rare, and so highly prized. Big data predominantly measures actions and outcomes rather than the psychological constructs (attitudes, beliefs, perceptions, dispositions etc.) that are foundational to many streams of individual-level IS research. Attempts to develop models from big data may shift researchers' focus of theorizing away from psychological models, and may reduce our ability to make strong inferences about cognitive antecedents in theoretical models. Some expansion of focus is no doubt a good thing, but if we as a field value our tradition of psychologically-based theories, it seems important that we take steps to link big data research to those theories that have traditionally been a source of strength for our discipline.

Because it is impossible to infer motivations from actions, researchers might make unsubstantiated guesses about intrapersonal antecedents or moderators; over time, this risks splitting off from an important body of IS theory. The value of triangulating big data with small data, and primary and secondary data becomes obvious (George et al., 2014). We therefore recommend that researchers may consider conducting multi-phase studies that survey subsets of subjects in big datasets to validate associations between psychological antecedents and actions or outcomes measured in big datasets. Once these connection points are established, subsequent research phases could focus on theoretical models of associations found between big data behaviors and outcomes, confident in the links backwards to intrapersonal constructs.

6.3.2. No research project is an island

Building theory is a long-term process of accumulating knowledge generated via many different research efforts. Researchers must be especially conscious of the need to fit in and contribute to an ongoing research stream, as it may be easy to slip into the organization's mindset and get excited about novel findings that do not inform theory. By itself, a novel empirical result may be highly relevant and of great interest, but to contribute to the process of building theory, an empirical claim is not enough. Researchers who advance new knowledge claims that explain why things happened (not just what happened) and that are situated in an ongoing theoretical conversation stand to make the greatest contribution to IS research.

6.3.3. Thoroughly explore boundary conditions

An important element of building theory is articulating boundary conditions that establish the limits beyond which hypothesized relationships are unlikely to hold (Gray & Cooper, 2010). For example, a theory may predict (a) an effect on some outcomes but not others, (b) an effect in some contexts but not others, or (c) a stronger or weaker effect under certain moderating conditions. When sample sizes are small, as if often the case when using the hypothetico-deductive approach, there is less statistical power to test theoretical boundaries. Theories whose boundaries are not tested may leave readers with an overly optimistic impression of how generalizable they are. However, when leveraging big data, there is a greater incentive to move from broad general claims about direct effects towards more nuanced models specifying the theoretical conditions under which a theory holds and where it does not, because researchers are less limited by sample size constraints. Big data therefore provides an ideal opportunity for researchers to test theoretical boundary conditions that would not be possible in small samples. However, in some contexts there may be no theoretical explanations for boundary conditions, but post-hoc exploratory analysis may reveal their existence. In such a case, we recommend that researchers perform (and report) empirical tests of potential moderators in order thoroughly explore the conditions under which a theory holds and does not hold in a single big dataset. Further, because big data is often longitudinal in nature, many of the shortcomings of cross-sectional design can be overcome through research methods such as panel data design or process-based models (c.f., Mackenzie, 2000). The presence of time-based and other empirical boundary conditions is extremely important both to future theorists and for practice. Researchers contribute to cumulative knowledge building by publishing such post-hoc, exploratory analyses that delineate the situations under which support for their theory is stronger and weaker, where it holds and where it does not hold.

7. Conclusion

As we move from data scarcity to data abundance, to leverage big data there is a need to re-visit and modify—not discard—traditional approaches to IS research. Our academic endeavors, whether qualitatively or quantitatively inclined, are at least as much about sensemaking and abstracting knowledge into generalizable building blocks as they are about identifying empirical patterns. High-quality IS research requires not only analyzing "*what* is happening?" in a salient IT-related phenomenon but also identifying "*why* is it happening?" As readers of this journal can attest, explaining the "*why*"—through building and testing theory—remains of central importance to our field. While we agree with many of the optimistic assessments of Agarwal and Dhar (2014) and Goes (2014) about the potential of big data to enhance our discipline, we also caution that an uncritical embrace of big data empiricism will do little to generate either durable, cumulative knowledge or ground-breaking conceptual insights. Further, we reject the idea that big data means that "algorithms find the patterns and the hypothesis follows from the data. The analyst doesn't even have to bother proposing a hypothesis any more" (Steadman, 2013). Instead, we believe the traditional objectives of building and testing IS theory can be enriched with big data. Our goal has been to continue the dialogue started by Agarwal and Dhar (2014), Rai (2016), Goes (2014), and others about how we might conduct our practice of theory building research differently in the face of the big data deluge.

We hasten to note that some researchers may have instinctively adopted similar approaches to theory-informed big data research. Our contribution is to go beyond prior publications in describing a clear rationale for methodological choices and to begin the process of codifying best practices. We believe there is value in articulating our hybrid model and associated implications in order to help those who would like to engage big data in theory building research but are unsure how. More generally, we hope that this paper also serves to focus attention on the underlying opportunities and issues, and that others might advance our thinking with even better approaches for reflective inclusion of big data into the practice of IS research. While there has been some discourse on specific statistical methods surrounding big data (e.g., Shmueli, 2010), to date we have seen no good methodological discussion in IS journals or related management disciplines, and so perhaps this is also an opportunity for IS researchers to lead, and where other disciplines may follow.

There is certainly room for multiple approaches to IS research, and those approaches will coexist better if each is judged by the standards to which it aspires. Articulating clearly the approach adopted for a given project will help reviewers and readers assess a manuscript's claims against relevant objectives and standards. Our goal is to encourage IS researchers to leverage big data through a hybrid approach that fits in the rich IS tradition of building and testing theories with managerial, organizational, and societal implications, which "sets us apart from practitioners and consultants" (Gregor, 2006, p. 613). We are optimistic that IS researchers can incorporate big data into our research traditions. To aid in this transition we summarize key implications in Table 4.

These implications also offer a perspective for journal reviewers assessing the contributions of IS research inspired by the availability of big data. For instance, as we articulated above, strong conceptual claims require close alignment of phenomenon, theoretically-informed research questions, and data; reviewers should look for evidence of this tight alignment. In general, we hope that reviewers can use each of our implications to inform their assessment of how well a manuscript balances the opportunities of data availability with the objectives of cumulative knowledge generation.

The shift from data scarcity to data abundance is profoundly changing how organizations make decisions and generate value, and has already impacted IS research in a variety of ways, with much more surely to come. In these exciting (and perhaps anxious) times, it is important to see the potential in big data to help improve theory development.

Our hybrid approach represents *one* way to develop IS theory, and certainly is *not* offered as a replacement for any other form of research. There remains a strong need for both traditional hypothetico-deductive and emergent big data empiricism approaches, as well as the plethora of other valuable approaches that are regularly featured in our journals. We believe that the practically-grounded sort of theory that may come from using our approach could be especially valuable by producing both conceptual and pragmatic knowledge claims. We hope that by re-sequencing the research process and by emphasizing different activities within key phases, our approach will encourage and guide IS researchers in making the most of this terrific new opportunity.

Acknowledgements

This research received support from the McIntire School of Commerce and the Center for Business Analytics at the University of

Table 4
Recommendations for theory-informed big data research.

Research phase	Recommendations
Project ideation	1. Drive your project based on theory-informed research questions, not convenience of data. 2. Know your data: immerse deeply in the phenomenon to identify what is unique and what is common. 3. Be prepared to walk away. Data is increasingly abundant. Be picky. 4. When big data is the best source of evidence, include data science expertise on your research team.
Gathering evidence	5. Validate your data: it may not mean what you think it means. 6. Thoughtful iteration is your friend: if you change a measure, revisit your constructs to make sure they still match up. 7. Research your data: big data is longitudinal. This long-term history is a blessing and a curse. 8. Plan ahead: document plans for analysis before you start. 9. Data can't think. Big data captures actions, not attitudes nor beliefs.
Asserting knowledge claims	10. Document your data: map the route all the way for data origin (by a person or system), through acquisition and preparation, to analyses. 11. Find the limits. Identify and describe boundary conditions. 12. Share practical empirical findings that emerge during data analysis. 13. Build cumulative knowledge. Situate conceptual knowledge claims within prior literature.

Virginia.

References

Abbasi, A., Sarker, S., & Chiang, R. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2), Retrieved from http://aisel.aisnet.org/jais/vol17/iss2/3.

Agarwal, R., & Dhar, V. (2014). Editorial—Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research, 25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546.

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired, 16*(07), Retrieved from http://www.wired.com/2008/06/pb-theory/.

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14*(4), 496–515. https://doi.org/10.5465/AMR.1989.4308374.

Berente, N., Seidel, S., & Safadi, H. (2019). Data-driven computationally-intensive theory development. *Information Systems Research*. https://doi.org/10.1287/isre.2018.0774 (Forthcoming).

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences.* Cambridge, Massachusetts: MIT Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for generalized causal inference.* (Houghton Mifflin).

Chan, J., & Ghose, A. (2014). Internet's dirty secret: Assessing the impact of online intermediaries on HIV transmission. *MIS Quarterly, 38*(4), 955–976.

Computer Sciences Corp (2012). Big data universe beginning to explode. Retrieved October 26, 2015, from http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode.

Davis, M. S. (1971). That's interesting. *Philosophy of the Social Sciences, 1*(2), 309.

Davis, G. F. (2015). Editorial essay what is organizational research for? *Administrative Science Quarterly, 60*(2), 179–188. https://doi.org/10.1177/0001839215585725.

Dibbern, J., Winkler, J., & Heinzl, A. (2008). Explaining variations in client extra costs between software projects offshored to India. *MIS Quarterly, 32*(2), 333–366. https://doi.org/10.2307/25148843.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460. https://doi.org/10.1511/2014.111.460.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal, 57*(2), 321–326. https://doi.org/10.5465/amj.2014.4002.

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forum and search data for sales prediction of high-involvement projects. *Management Information Systems Quarterly, 41*(1), 65–82.

Goes, P. (2014). Editor's comments: big data and IS research. *Management Information Systems Quarterly, 38*(3), iii–viii.

Gray, P. H., & Cooper, W. H. (2010). Pursuing failure. *Organizational Research Methods, 13*(4), 620–643. https://doi.org/10.1177/1094428109356114.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611–642.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: A cautionary note on the potential for bias. *Clinical and Translational Science, 7*(4), 342–346. https://doi.org/10.1111/cts.12178.

Kitchin, R. (2014a). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), Retrieved from http://bds.sagepub.com/content/1/1/2053951714528481.short.

Kitchin, R. (2014b). *The data revolution: Big data, open data, data infrastructures and their consequences.* Sage.

Lee, A. S. (1989). A scientific methodology for MIS case studies. *MIS Quarterly, 13*(1), 33–50. https://doi.org/10.2307/248698.

Lee, A. S. (1991). Integrating positivist and interpretive approaches to organizational research. *Organization Science, 2*(4), 342–365.

Lewin, K. (1945). The research center for group dynamics at Massachusetts Institute of Technology. *Sociometry, 8*, 126–135.

Mackenzie, K. D. (2000). Processes and their frameworks. *Management Science, 46*(1), 110–125.

McKenna, B., Myers, M. D., & Newman, M. (2017). Social media in qualitative research: Challenges and recommendations. *Information and Organization, 27*(2), 87–99. https://doi.org/10.1016/j.infoandorg.2017.03.001.

Müller, O., Junglas, I., vom Brocke, J., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: Challenges, promises and guidelines. *European Journal of Information Systems*. https://doi.org/10.1057/ejis.2016.2 Retrieved from.

Nelson, L. K. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research,* 1–40.

Pentland, B. T., Pentland, A. P., & Calantone, R. J. (2017). Bracketing off the actors: Towards an action-centric research agenda. *Information and Organization, 27*(3), 137–143. https://doi.org/10.1016/j.infoandorg.2017.06.001.

Prensky, M. (2009). *H. sapiens* digital: From digital immigrants and digital natives to digital wisdom. *Innovate: Journal of Online Education, 5*(3), 1.

Rai, A. (2016). Editor's comments. *MIS Quarterly, 40*(1), iii–x.

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age.* Princeton University Press.

Schutz, A. (1962). Common-sense and scientific interpretation of human action. In M. Natanson (Ed.). *Collected papers I* (pp. 3–47). Netherlands: Springer. https://doi.org/10.1007/978-94-010-2851-6_1.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference. vol. xxi.* Boston, MA, US: Houghton, Mifflin and Company.

Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: A research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems, 23*(4), 433–441.

Shi, Z., Lee, G., & Whinston, A. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. *Management Information Systems Quarterly, 40*(4), 1035–1056.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330.

Starbuck, W. H. (2016). 60th anniversary essay how journals could improve research practices in social science. *Administrative Science Quarterly,* 165–183. https://doi.org/10.1177/0001839216629644.

Steadman, I. (2013, January 25). Big data and the death of the theorist. Retrieved July 14, 2016, from http://www.wired.co.uk/article/big-data-end-of-theory.

Veiga, J. F., Keupp, M. M., Floyd, S. W., & Kellermanns, F. W. (2014). The longitudinal impact of enterprise system users' pre-adoption expectations and organizational support on post-adoption proficient usage. *European Journal of Information Systems, 23*(6), 691–707.

Venkatesh, V., Brown, S. A., & Sullivan, Y. W. (2016). Guidelines for conducting mixed methods research: An extension and illustration. *Journal of the Association for Information Systems, 17*(7), Retrieved from http://aisel.aisnet.org/jais/vol17/iss7/2.

Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems, 4*(2), 74–81.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences. vol. 111.* Rand McNally Chicago.

Yin, R. K. (1994). *Case study research: Design and methods.* Sage Publications.