Journal of the Association for Information Systems (2018) **19**(12), 1253-1273
doi: 10.17705/1jais.00526

EDITORIAL

ISSN 1536-9323

Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research

Wolfgang Maass[1], Jeffrey Parsons[2], Sandeep Purao[3],Veda C. Storey[4], Carson Woo[5]

[1]Saarland University, Germany, wolfgang.maass@iss.uni-saarland.de
[2]Memorial University of Newfoundland, Canada, jeffreyp@mun.ca
[3]Bentley University, U.S.A., spurao@bentley.edu
[4]Georgia State University, U.S.A. VStorey@gsu.edu
[4]The University of British Columbia, Canada, carson.woo@sauder.ubc.ca

Abstract

The era of big data provides many opportunities for conducting impactful research from both data-driven and theory-driven perspectives. However, data-driven and theory-driven research have progressed somewhat independently. In this paper, we develop a framework that articulates important differences between these two perspectives and propose a role for information systems research at their intersection. The framework presents a set of pathways that combine the data-driven and theory-driven perspectives. From these pathways, we derive a set of challenges, and show how they can be addressed by research in information systems. By doing so, we identify an important role that information systems research can play in advancing both data-driven and theory-driven research in the era of big data.

Keywords: Big Data, Data Analytics, Data-Driven Research, Theory-Driven Research, Abstraction, Generalization, Systems Analysis, Requirements, Information Systems Research

Suprateek Sarker was the accepting editor.

1 Introduction

Expectations remain high for the potential of big data to advance our understanding of business, society, and science (Baesens, Bapna, Marsden, Vanthienen, & Zhao, 2016; Bell, Hey, & Szalay, 2009; Dhar, 2013; Goes, 2014; Günther, Mehrizi, Huysman, & Feldberg, 2017; Gupta, Deokar, Iyer, & Sharda, 2018; Maass et al., 2017; Mayer-Schonberger and Cukier, 2013; Markus & Topi, 2015). Information systems (IS) scholars have analyzed various issues in advancing big data research. For example, Abbasi, Sarkar, & Chiang (2016) propose a big data research agenda following behavioral, design or economics research approaches, building on the idea of the information value chain (i.e., data, information, knowledge, decision, and actions). Rai (2016) provides insights on the role of theory and suggests that synergies between big data and theory are yet to be realized.

To better understand research opportunities using big data, we distinguish two perspectives: data-driven research and theory-driven research. Data-driven research is an exploratory approach that analyzes data to extract scientifically interesting insights (e.g., patterns) by applying analytical techniques and modes of reasoning. Theory-driven research is a more traditional approach of conducting scientific inquiry that starts with developing hypotheses, followed by collecting and analyzing data to test these hypotheses and drawing theoretical conclusions based on the results. Scholars have recognized that the data-driven and theory-driven research perspectives should be mutually reinforcing in the era of big data (e.g., Siegfried, 2013; West, 2013; Kitchin, 2014).

The objective of this paper is to examine data-driven and theory-driven perspectives for conducting research and, by doing so, identify opportunities and challenges that information systems (IS) researchers can, and should, respond to in the big data era. To achieve this objective, we first propose a framework for identifying mutually reinforcing interactions between data-driven and theory-driven research efforts. The framework is then used to specify roles IS researchers can play at the intersection of the two perspectives. The roles are presented in terms of four challenges, the resolution of which can be assisted using methods and techniques developed in the IS discipline. The contributions of this paper are twofold: (1) a framework for conducting research in the big data era by combining data-driven and theory-driven perspectives, and (2) proposed ways in which IS researchers can undertake this work.

The next section reviews data-driven versus theory-driven research perspectives. This is followed by our proposed framework for IS research in the big data era. Based on this framework, we identify four challenges for IS researchers and offer suggestions for their resolution. Finally, the conclusion reinforces the continued need for IS researchers to play a central role in research at the intersection of data-driven and theory-driven perspectives.

2 Two Research Perspectives: Data-Driven and Theory-Driven

Data-driven research uses exploratory approaches to analyze big data to extract scientifically interesting insights (Kitchin, 2014). Due to the complexity of the environments and processes that generate data, there may not be a strong theoretical base for the questions being studied. Data-driven research is typically described in terms of the following tasks, which may require iteration (Jagadish, 2015; Shmueli & Koppius, 2011):

(1) identifying research question(s) based on a knowledge gap in a domain of interest;

(2) creating/obtaining sources of data germane to relevant phenomena in the domain;

(3) cleansing, extracting, annotating data streams to prepare for analyses;

(4) integrating, aggregating, and representing data to detect insights (e.g., correlations, patterns);

(5) analyzing and modeling data to place correlations and patterns in context; and

(6) interpreting the patterns to arrive at solutions and insights.

Data-driven research has been popular in some of the natural sciences, such as meteorology and astronomy, where large amounts of data are collected by sensors and other instruments (Sellars et al., 2013; Pankratius & Mattmann, 2014). This mode of science is considered effective, at least in part, because the size of the datasets is simply "big." The scale at which data are generated and used provides reliability that simply cannot be achieved with conventional scientific approaches. Although researchers may appeal to prior theory while interpreting their findings, this is often feasible only after the analysis. The primary contributions of data-driven research, then, are: (1) patterns extracted from the analysis of large data sets; and (2) insights derived from these patterns.

Data-driven research, as its name suggests, relies on the identification of patterns (robust correlations between sets of variables) to yield insights on empirically interesting phenomena based on the data available (rather than predicted based on theory). Because patterns are determined by relationships in the available data, scholars engaged in data-driven research face the challenge of building a cohesive body of knowledge about phenomena. Although interesting outcomes might be produced (similar to what early research found as unexpected correlations/associations (e.g., Bentley, O'Brien, & Brock, 2014; Davenport, Barth, & Bean, 2013; Dhar, 2013; Chan, Ghose, & Seamans, 2016), these results may not fit an existing theory of the domain, particularly during exploratory research on emerging topics.

In contrast, theory-driven research focuses on identifying abstract constructs and the relationships among them, and is usually described in terms of the following tasks (Andersen & Hepburn, 2016), which may include iterations:

(1) identifying a research gap;

(2) deriving research questions from existing or extended theory;

(3) formulating hypotheses to address the questions;

(4) designing studies to minimize confounding effects;

(5) collecting data using appropriate instruments; and

(6) analyzing data to draw inferences.

Theory-driven research has dominated the social and organizational sciences. A theory identifies constructs and relationships among them that are abstracted from specific phenomena. Over time, a theory codifies a body of knowledge about phenomena within its scope. Theories are often developed from deep reflection, sometimes aided by insights from small datasets (e.g., Eisenhardt, 1989). Traditionally, theory testing also uses relatively small datasets for several reasons. First, the cost of experimental design and data collection, including gaining access to these

complex phenomena, is high. Second, collecting data from every possible perspective can require a researcher to play conflicting roles. Theory-testing for these phenomena becomes a quest to discover statistical regularities in examined instances.

Empirical work in theory-driven research has historically been restricted due to demands on time, effort, and cost. The era of big data brings with it: (1) the ability to consider (close to) an entire population instead of a sample; (2) a lower cost of data acquisition (compared to traditional modes); and (3) the possibility of exploring many more correlations (on demand). The opportunity in the era of big data is not to make the scientific method obsolete (e.g., Anderson, 2008), but rather, to combine theory-driven and data-driven research to realize the potential to transform how research is conducted in the social and organizational sciences. Many of today's big problems (e.g., developing smart cities (Batty, 2013), solving poverty, and addressing climate change (Hampton et al., 2013)) require multidisciplinary solution approaches that combine the power of a data-driven approach with the deep domain understanding provided by domain theories.

3 An Information Systems Framework for Research in the Era of Big Data

This section develops a framework for information systems research in the era of big data, in which data-driven and theory-driven perspectives are combined.

3.1 The Two Perspectives

The differences between the data-driven and theory-driven perspectives can be problematic. An exclusive emphasis on big data analytics, without considering domain theory, can lead to the identification of correlations, trends, and patterns that provide answers to situated questions, but might not contribute to enduring scientific knowledge. Conversely, an exclusive emphasis on domain theory with continued use of small datasets (often collected at high costs from primary sources) might result in missed opportunities to make important discoveries using big data. Figure 1 depicts the two perspectives as alternative approaches to conducting research, represented as solitudes in which work is carried out independently.

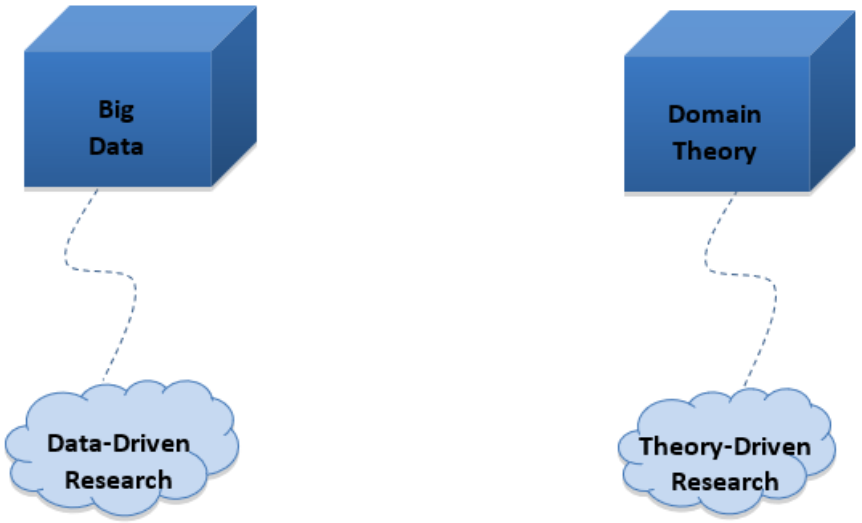


Figure 1. Two Perspectives: Data-Driven and Theory-Driven Research

3.2 Examples from the Information Systems Literature

Although work in each perspective can lead to important insights, emphasizing a single perspective can also lead to the loss of opportunities. To illustrate, consider two studies from the IS literature that illustrate the *data-driven* and *theory-driven* perspectives, respectively. (We revisit these studies later to show the benefits, in each case, of considering the other perspective.)

Greenwood and Agarwal (2016) studied the temporal relationship between the introduction of Craigslist to various urban areas in the United States (specifically, in Florida) and subsequent increases in the reported cases of asymptomatic HIV diagnosed by hospitals in the region. The authors found that cases of HIV increased after the introduction of Craigslist. Differences were observed in the strength of the effect for various racial groups, gender, and socioeconomic status. Estimates of the economic cost were provided.

This paper falls in the category of *data-driven research* for several reasons. First, key to the analysis

is the combining of two independent data sources: (1) hospital admission and diagnosis data; along with (2) Craigslist data about introduction of the service to various areas. Second, the hospital admission dataset consists of records for approximately 12 million patients. Finally, the paper focuses extensively on the datasets, the data analysis and the practical implications of the findings, rather than on theory development to justify either the choice of datasets or hypotheses development. The research is motivated by prior work on the effects of matching platforms on engagement in risky behavior, but does not focus on abstract construct development or the identification of causal relationships among constructs. Instead, it focuses on the effects of such platforms on reducing transaction costs and information asymmetry, where the authors provide insights following the extraction of associations and patterns.

The second example is Xiao and Benbasat's (2015) study of product recommendation agents in electronic commerce, which focuses on how the design of warning messages can facilitate the detection of bias in recommendations. The authors found that the effectiveness of warning messages depends on whether the warnings are accompanied with advice on how to check for bias and whether that advice is framed positively or negatively.

This study fits the category of *theory-driven* research for several reasons. First, the research is extensively motivated by signal detection theory, which accounts for phenomena in decision-making in uncertain contexts, in which signals must be extracted from available information to guide decisions. The authors extend this theory to the context of recommendation agents by considering how the design of warning messages (rather than just their presence or absence) contributes to the ability to detect recommendation bias. Thus, they contextualize signal detection theory within the specific case of making sense of online recommendations. Second, although the theoretical propositions are tested in an online experiment, the design is based on a small sample size as traditionally associated with experimental studies. Such studies have limited variation in the design space for the manipulation of independent variables.

3.3 Paths to Connect the Perspectives

To connect the perspectives, two important pathways are proposed, as shown in Figure 2. The path from left to right (top arrow) represents the possibility of progressing from patterns extracted during big data analytics to the abstraction and generalization needed for domain theory development and refinement. The path from right to left (bottom arrow) captures the importance of identifying the data sources and types of analyses needed for theory testing and refinement. These paths may be manifested in various ways, as illustrated in the examples below.

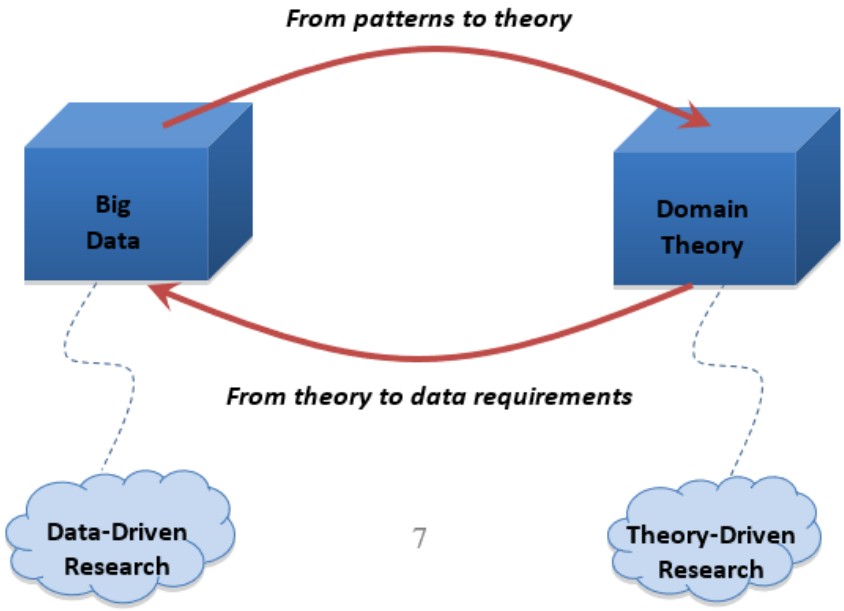


Figure 2. Paths to Connect Data-Driven and Theory-Driven Research

3.3.1 From Patterns to Theory

The path from Big Data to Domain Theory (Figure 2) starts with data-driven research, which focuses on identifying patterns that represent relationships among concepts. These patterns can be further analyzed in at least two ways. First, patterns extracted from big data can be used to derive insights about

domain theories. Second, in attempting to interpret extracted patterns in terms of existing theory, relationships can be exposed as potentially spurious if theory suggests that other factors can account for observed relationships (Bentley et al., 2014).

As researchers engage in data-driven research, they perform specific tasks, including data preparation, exploratory analytics, choice of variables, and model selection (Shmueli & Koppius, 2011). These tasks can use a wide range of techniques and algorithms. For example, supervised learning fits input data to given output data assumed as being "ground truth" and, thus, implicitly learns relationships between variables (aka features). Unsupervised learning does not leverage true answers but searches for joint probability density functions of input data that indicates some intrinsic structure (e.g., principal component analysis or cluster analysis) (Hastie, Tibshirani, & Friedman, 2009). Reinforcement learning (RL) lies in between by optimizing the selection of actions in a given state according to a reward function without learning from correct actions (Sutton & Barto, 1998).

Regardless of the specific algorithms and techniques used, patterns are recognized implicitly. From a researcher's perspective, systems are black boxes that transform input data into output data, based on a specific quality. From the implicitly learned patterns, predictions, prescriptions, and classifications can be made. Research has attempted to understand black boxes by identifying the internal structures responsible for predictions (Vidovic et al., 2015). This is considered to be a first step towards interpreting implicit patterns. With black-box models, researchers can study behavior and extract descriptions of explicit patterns. For instance, the Go world-champion Lee Sedol studied moves taken by the system AlphaGo to extract explicit patterns that enabled him to obtain an unusual 22-game winning streak against human opponents. (Economist, 2017). Subsequently AlphaGo zero found behavioral patterns by applying reinforcement learning that easily defeated AlphaGo (Silver et al. 2017).

Extracted explicit patterns may be interpreted or explained to arrive at new scientific insights. This requires placing explicit patterns within the context of a domain to develop, support, refute, or refine constructs from an underlying theory (which may require iteration). For example, classification of patient data can reveal (with a given level of certainty) predictions on cancer susceptibility, recurrence and mortality (Cruz & Wishart, 2006).

Another example of data-driven research is a study that integrates and analyzes weight data from 2416 population-based studies on 128.9 million participants (NCD Risk Factor Collaboration, 2016). Weight categories are defined by standard deviations from medians. The results are interpreted by geographical regions and show, for instance, the decrease of moderately and severely underweight girls in India and the prevalence of obesity on various Polynesian islands (NCD Risk Factor Collaboration, 2016). This study is a prime example of research with big data and small theories. The data is used to derive linear and nonlinear body-mass index (BMI) trends in geographical regions based on a Bayesian hierarchical model. The study concludes that age-standardized BMI is increasing worldwide, thus providing a basis for studying social, psychological, and medical questions that explain local and global BMI increases by theories only partially developed today (Finucane, Paciorek, Danaei, & Ezzati, 2014).

A second manifestation of the path from patterns to theory occurs when observed patterns fail to account for factors (theoretical or otherwise) that are missing in the available data, but would better explain observed relationships, rendering particular observed patterns spurious. For example, in early analysis Google Flu Trends (GFT) identified a relationship between specific terms used in Google searches in a region and incidence of influenza, with the objective of predicting the prevalence of influenza simply based on an analysis of Google searches. While this approach worked initially, it later "failed miserably" (Lazer & Kennedy, 2015). In 2013, for example, GFT predicted twice as many cases as reported by the Centers for Disease Control and Prediction (CDC), even though it was constructed to predict these numbers. Later analysis showed that part of the issue was the correlation between the search terms used and winter, which is when influenza is most prevalent. As Lazer, Kennedy, King, & Vespignani (2014) note: "the initial version of GFT was part flu detector, part winter detector" (p. 1203). This example illustrates a situation where a big data prediction ended up being incorrect and domain theory helped to detect the cause. Here, domain theory helped reveal the accuracy of the results by triangulating the patterns.

This example shows how results obtained from data analysis enable domain theorists to go beyond the abstractions offered by data-driven researchers to add explanations and interpretations that map the meanings of explicit patterns against constructs and relationships that are part of domain theories. Figure 3 shows how this can be achieved by aligning results from big data analytics with the constructs and relationships (existing or new) in domain theories.

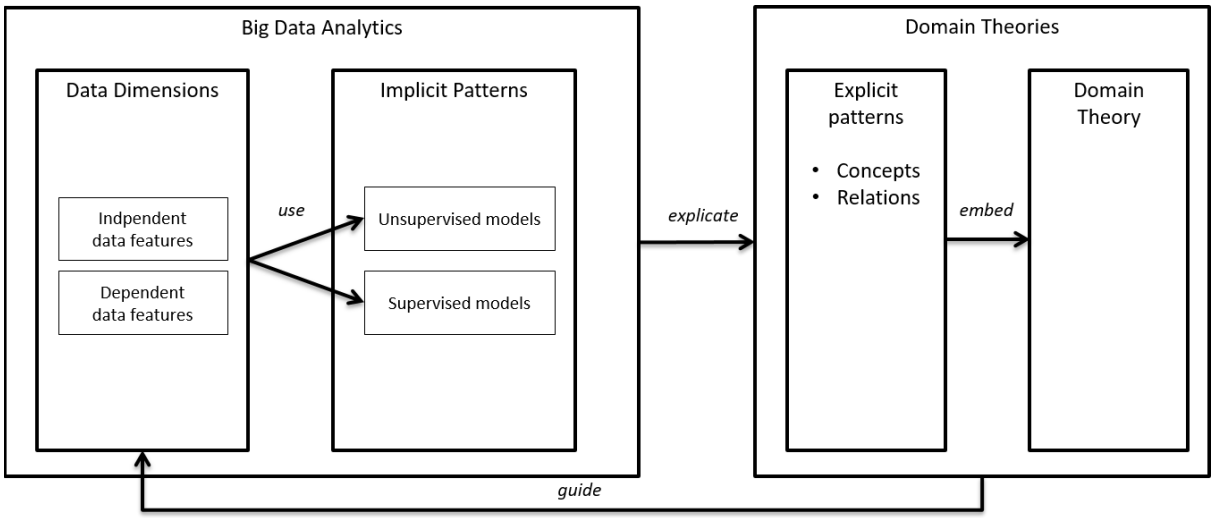


Figure 3. Mapping Big Data Analytics to Domain Theory

To illustrate the path from data-driven to theory-driven research, consider again the work of Greenwood and Agarwal (2016). The authors showed that the introduction of Craigslist in regions of Florida was followed by increased incidence of HIV diagnosis at hospitals in these regions. Given the data used, these results are specific to a particular technology and a particular health outcome in a particular geographic region.

One way such work can contribute to theory development is to abstract beyond the particular context (casual sexual encounters organized via Craigslist and incidence of HIV) to more general concepts (e.g., personal interactions resulting from online connections and associated health consequences) from domain theories. Greenwood and Agarwal (2016) do speculate on plausible mechanisms towards such abstractions. In addition, by drawing on behavioral theory, it may be possible to understand, at a general level, what triggers users of such services to engage in certain behaviors in real life. Such theorizing could benefit from interaction with domain specialists from the health care setting (e.g., epidemiologists), as well as from interaction with psychologists, to understand the general factors that determine the extent and ways in which individuals choose to engage in behaviors with known risks. Then, design science researchers could contribute to understanding how design features of platforms might contribute to or be used to mitigate such behaviors.

The Patterns to Theory path (Figure 2) does not diminish the importance of the findings from data-driven research—including how the identification of patterns and clusters, as well as the initial abstractions from these—yield insights for decision makers. Rather, this path points out opportunities for extending the interpretations of these findings by considering specific constructs and relationships from appropriate domain theories. Relating the results obtained by data-driven research to constructs and relationships in domain theories might even lead to the identification of new constructs or relationships that could enrich or refine existing domain theories.

3.3.2 From Theory to Data Requirements

The path from theory-driven to data-driven research shows that domain theory can guide the search for patterns by identifying possible constructs and relationships that can be used in the analysis. This can lead to collaboration across the two perspectives, thus contributing to data analytics.

Theories express abstract concepts and relationships among them. To test hypotheses regarding relationships among constructs, the latter are operationalized. As part of the scientific process, operationalization is employed when a construct, which is not directly measurable, is characterized by one or more measurable variables that act as a surrogate for the construct. This makes it possible to specify manipulations (in the case of experiments) and define measurement items (in experimental and survey research). Testing domain theories using big data, however, is more complex, because it has additional challenges including identification of the scope, source, and quality of the data. Domain theorists may suggest: (1) new data sources that data analytics researchers may not have considered, or (2) novel combinations of data sources. In addition, the availability of big data sources with many variables can provide a useful setting to identify conditions that specify the boundaries of existing theories. This can be done by using the (big) data to identify additional conditions under which the predictions do or do not hold, thus aiding in overcoming the challenge of reproducibility in behavioral research (Open Science Collaboration, 2015).

To illustrate the path from Domain Theory to Big Data (Figure 2), consider again the work by Xiao and Benbasat (2015). The authors theorized about the effects of warning messages (absence or presence, role of advice, and framing as negative and positive) on the detection of bias in recommendations. A key aspect of that study is using theory to guide the design of warning messages. However, the traditional data collection approach severely limits the way in which design features (such as the form of advice) can be manifested in an IT artifact. Indeed, the authors note: "Given the multitude of means by which bias can be introduced by PRAs, future research is needed to explore what the most appropriate informational content of risk-handling advice is for biases introduced from various other sources" (Xiao & Benbasat, 2015, p. 809). It is now possible to run large-scale field experiments in which many variations on independent variables of interest can be simultaneously tested by assigning users in natural settings randomly to a broader range of conditions that systematically vary factors of interest (Lukyanenko, Samuel, & Parsons, 2018). Thus, theory can be used to guide the design of data collection on a broad range of variables. Figure 4 illustrates mapping from domain theories to big data.

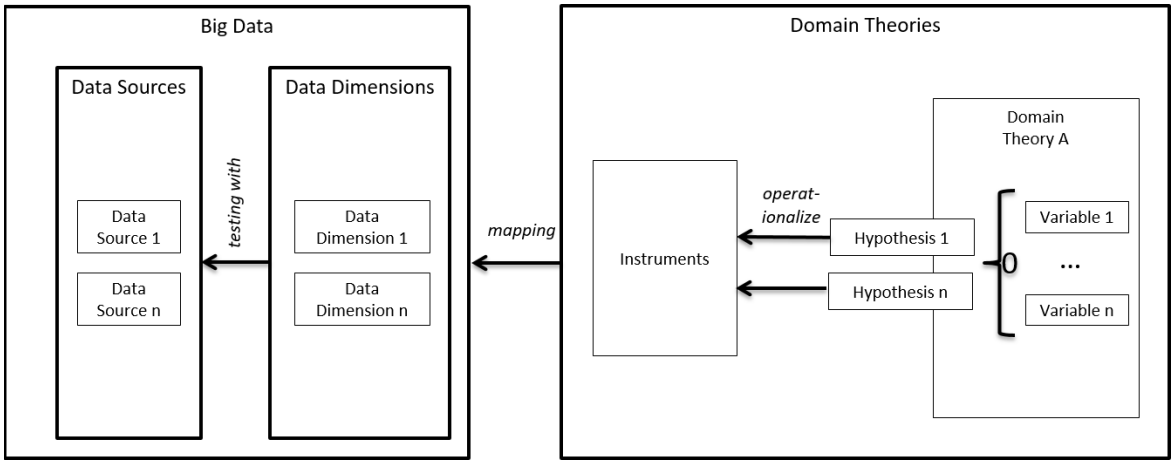


Figure 4. Mapping Domain Theory to Big Data Analytics

Another way the path from theory to data requirements can be manifested is by conditioning data requirements based on ethical considerations. For example, Dressel and Farid (2018) showed that a complex risk assessment model (with 137 features) used to predict criminal recidivism performed no better than a simple linear model with two features. The model also did not perform any better than novices (recruited via Amazon Mechanical Turk) who were provided with three indicators (sex, age, and criminal history). However, the complex model showed a higher level of bias based on race—tending to overpredict recidivism rates for black offenders and underpredict recidivism rates for white offenders. In another application domain, the Tay chatbot was launched by Microsoft in 2016 to embed machine learning as a way to engage in realistic conversations with Twitter users. It was quickly shut down after other Twitter users, who were engaged with Tay, trained Tay to generate offensive and abusive tweets (Neff & Nagy, 2016). Work such as this reinforces the importance of variable or feature selection in preparing for data mining and exposes limitations of methods based solely on data analytics. Such variable selection can be guided both by domain theory and overarching ethical principles intended to embed fairness in the resulting models.

3.4 A Framework for Information Systems Research at the Intersection

We now propose a framework that identifies potential information systems research challenges and opportunities at the intersection of data-driven and theory-driven approaches. An important component of the framework is the research knowledge that can bridge the two perspectives. The paths identified in Figure 2: "From patterns to theory" and "From theory to data requirements" require researchers to engage in tasks for which no single individual may be perfectly equipped because they require knowledge and skills related to data analytic techniques, as well as expertise in relevant domain theories. Hence, due to the complexity and heterogeneity of research knowledge on both sides, the need for interaction and bridging of the two arises, as shown in Figure 5. Information systems researchers have knowledge in areas such as systems analysis and design (Avison & Fitzgerald, 2003), and fulfill a liaison role (Mathiassen & Purao, 2002) between diverse stakeholders that may be required to bridge theory-driven and data-driven perspectives.

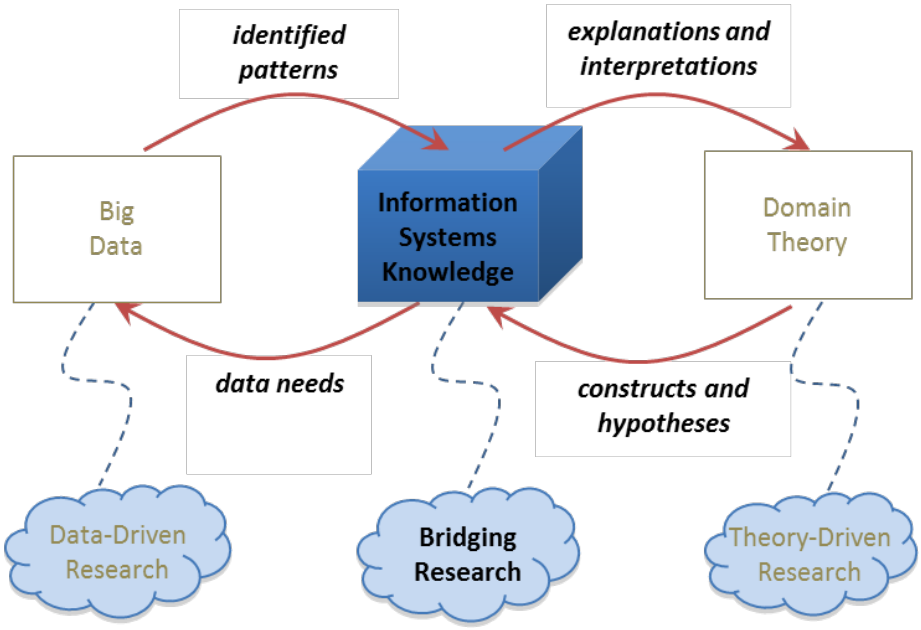


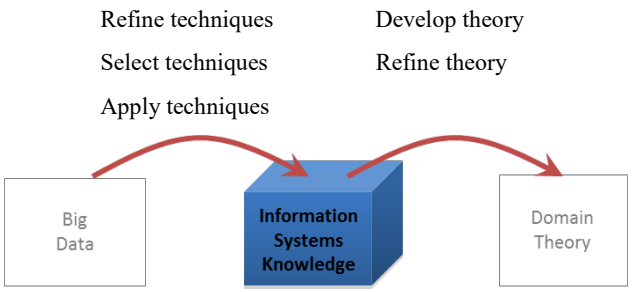
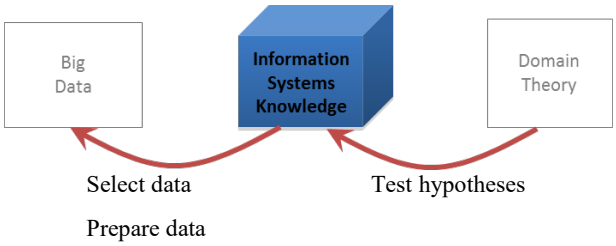
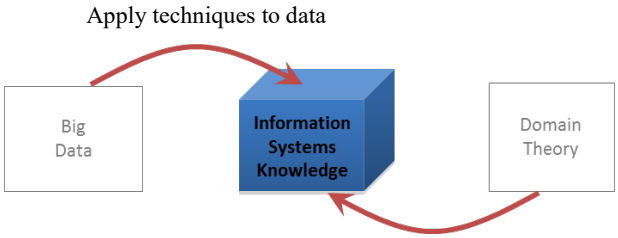
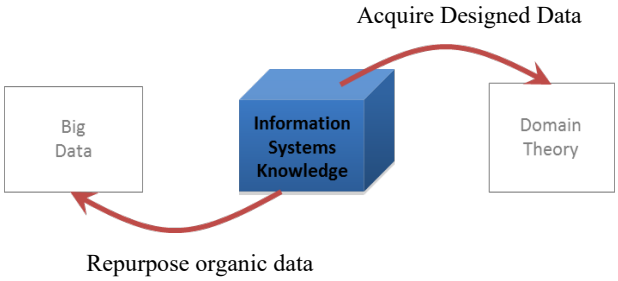
Figure 5. Framework for IS Research at the Intersection of Data-Driven and Theory-Driven Perspectives

The required knowledge at the intersection consists of tools to engage in two bridging tasks: (1) synchronizing research activities in data-driven and theory-driven research; and (2) applying methods that work on results from and assign requests to both sides. On one hand, this bridging knowledge supports abstraction and generalization tasks on identified patterns derived by big data analytics for incorporating results into domain-specific explanations and interpretations. On the other hand, it supports transforming domain-driven hypotheses into specifications of what data is needed for specific analytic techniques.

4 Research Challenges

By analyzing pairs of arrows in the framework (Figure 5), four challenges emerge that need to be resolved to conduct research effectively at the intersection of the data-driven and theory-driven perspectives. The possible interaction paths, with their associated challenges, are summarized in Table 1. The remaining pairs of arrows do not deal with interactions and, therefore, are not considered.

Table 1. Challenges at the Intersection of Data-Driven and Theory-Driven Research

Interaction paths	Challenge
	1. Reconciling competing approaches to creating or refining domain theories using big data
	2. Selecting data and analytic techniques to conduct theory testing
	3. Solving problems that are unsolvable from a single perspective
	4. Sharing data and models across research teams and projects

To realize synergies between data-driven and theory-driven research, the challenges identified in Table 1 must be resolved. These challenges are described below, with proposals for how they might be addressed by IS researchers.

4.1 Challenge 1: Reconciling Competing Approaches to Creating or Refining Domain Theories Using Big Data

When developing and refining domain theories using big data, knowledge creation approaches in data-driven and theory-driven research need to be reconciled, in the sense of making one consistent with the other. In data-driven research, the primary knowledge creation mechanism is identifying patterns by refining, selecting, and applying analytic methods to very large datasets. Researchers engaged in theory-driven research generate knowledge by developing and refining abstract constructs and relationships among them. Therefore, reconciling the two perspectives can be challenging. Researchers from both perspectives might find it difficult to specify what kinds of theories can be used to explain the analytical results.

4.1.1 Resolution by Abstraction and Generalization

Resolution requires understanding how general capabilities of data analytics and the availability of very large datasets can be used together to develop or refine domain theories. This entails abstracting and generalizing patterns identified from data analytics to support theory development and refinement. Abstraction (Woods & Rosales, 2010) involves hiding noise and details to focus on higher-level theoretical connections. Generalization (Parsons & Wand, 2013) involves the hierarchical organization of constructs to express levels of theoretical knowledge by making explicit subclass and superclass connections.

4.1.2 Example

Suppose patterns such as anxiety or informational uncertainty are identified from Twitter feeds. An original theory might suggest that anxiety and ambiguous information are key drivers of rumor-mongering, where anxiety is expressed through apprehensive statements. This theory can be refined based on data and patterns. For example, patterns extracted from the data could show that apprehension is insufficient. Instead, analysis might reveal that positive and negative emotional statements have different effects on rumor-mongering, thus providing a basis for refining the theory. The outcome might be an ontology of: emotional statements, authenticating statements, interrogatory statements, prudent disclaimer statements, belief/disbelief statements, and work statements (Baldoni, Baroglio, Patti, & Rena, 2012). Within the ontology, different kinds of statements could be hypothesized to produce different effects on rumors.

4.1.3 Role of Information Systems Research

IS research can address this challenge by providing modeling approaches that support explanation, interpretation, and generalization to express abstract, high-level concepts derived from patterns discovered using data analytics. The IS field uses a variety of modeling tools to express abstract concepts, including ontologies and conceptual modeling grammars (e.g., Weber, 2003). The abstraction of patterns (from data analytics) could allow us to identify, for example, new subclasses or to accommodate exceptional cases, resulting in theory refinement (Parsons & Wand, 2013). Alternatively, one might discover that patterns detected cannot be explained by an existing theory, thus necessitating new theory development. The difference between theory refinement and theory development rests on whether patterns discovered via data analytics can be accommodated within an existing theory.

An IS scholar can, therefore, play an *intermediary* role, acting as a liaison between the two perspectives to bring about agreement on how to understand the knowledge of both. In the above example, IS research on conceptual modeling (e.g., Wand, Monarchi, Parsons, & Woo, 1995) and ontology development (e.g., Sugumaran & Storey, 2002) can guide the design of an ontology of emotional statements, considering issues such as whether concepts need to be mutually exclusive and/or collectively exhaustive. In Table 1 (Row 1), the objective is to use information systems knowledge to connect research on the side where the arrows originate (data-driven) to research on the side where the arrows terminate (theory-driven). The skills that IS researchers can contribute are further elaborated using examples in Table 2.

Table 2. Roles of IS Researchers Responding to Challenge 1: Intermediary on Abstraction and Generalization

Research direction	What IS researchers can contribute
Support the expression of patterns	Apply, extend, and develop conceptual modeling approaches to represent, visualize, and communicate patterns (Woo, 2011)
Support the abstraction and interpretation of patterns into insights informed by theoretical concepts	Apply, extend, and develop techniques to map and interpret patterns into domain ontologies
Support analysis of big data that includes time-varying data	Extend work on data collection, representation, and use, which may include both linear and nonlinear analysis of time-varying data and other techniques (e.g. Chong, Han, & Park, 2017)

4.2 Challenge 2: Selecting Data and Analytic Techniques to Test Theory

Understanding what data are available and how to use that data are vital for theory testing using big data. Theory-driven research starts by generating hypotheses before proceeding to select and prepare the data to test the hypotheses. It operationalizes constructs and identifies how to manipulate and/or measure them, typically on a small scale. The data collected is examined for validity, reliability, and adequacy of sample size before being used to test hypotheses.

Theory testing traditionally does not consider important challenges in the analysis of large datasets, such as the heterogeneity of data sources and variations in the granularity of the data. This creates

difficulties in realizing the potential to use large datasets for theory testing and refinement. One important issue is whether and how domain theories can be used to determine how available data can be prepared for testing theory, and what additional data might be needed.

4.2.1 Resolution by Gathering and Assessing Requirements

This challenge can be resolved by determining data needs from theory-driven research. This is a form of requirements gathering in which the objective is to identify the data needed to test the theory. When some of the required data cannot be obtained using existing big data sources, the available data must be assessed or evaluated, based on appropriate criteria, so the most appropriate data can be selected. One might also discover that the data needed for testing is not available, requiring new data collection.

4.2.2 Example

Researchers have studied how cognitive maps can be used for understanding social and geographic environments (Unger & Wandersman, 1985). Spatial reference systems for investigating spatial knowledge (Blouin et al., 1993; Golledge, 1999), for example, have been evaluated in laboratory experiments (Shelton & McNamara, 2001) using relatively small datasets. With the potential for millions of self-driving cars, there is an unparalleled opportunity for a more realistic (authentic) evaluation of spatial reference systems using the large volume of data generated by in-car sensors. Information systems research on requirements analysis can be useful in identifying which data generated by sensors support the conceptualization and representation of spatial knowledge. From this, it might be concluded that visual sensors could be attached to self-driving cars to collect data that can be used for comparing users' reported perceptions versus sensor data. In this manner, self-driving cars become laboratories on wheels.

4.2.3 Role of Information Systems Research

Information systems research has produced systems analysis and design practices that can be useful in both authentic evaluation and new data collection. The goal in authentic evaluation is first to derive data requirements based on the theory to be tested, and then to select relevant data sources or manipulate available data sources to derive composites that map to data requirements. In systems analysis and design, requirements analysis techniques (Kotonya & Sommerville, 1998) guide the selection or design of information systems. These requirements techniques can be adapted to match constructs from theory with data sources appropriate to test it. However, big data can have unstructured, multiple representations and lack integrity due to its organic nature. IS research—such as work on assessing data quality (Wang, Storey, & Firth, 1995; Wang & Strong, 1996; Madnick & Zhu, 2006), understanding data semantics, and integrating data (Wand, Storey, & Weber, 1999)—can also be useful in authentic evaluation. As with the previous challenge, IS researchers can here play an *intermediary* role in linking the two perspectives. The specific roles IS researchers can play are further elaborated in Table 3.

Table 3. Roles of IS Researchers Responding to Challenge 2: Selecting Data and Analytic Techniques to Test Theory

Research direction	What IS researchers can contribute
Support the matching of constructs from data sources to domain theory	Apply, extend, and develop reverse engineering methods to map variables from data sources into constructs from domain theory (Chiang, Barron, & Storey, 1994)
Support the assessment and representation of data quality for multiple, heterogeneous data sources	Develop approaches to compute and represent quality metrics for heterogeneous data sources (Su, Huang, Wu, & Zhang, 2006)
Support the construction of composite variables from data sources	Develop techniques to evaluate potential composite variables for their usefulness by mapping them to on domain ontologies for interpretation
Support the representation of, and reasoning about, semantics implicit in data from multiple sources	Apply, extend, and develop techniques to represent and integrate the semantics of data from multiple sources (Evermann & Hallimi, 2008)

4.3 Challenge 3: Solving Problems that are Unsolvable from A Single Perspective

The challenge is how to use theory and data analytic techniques to solve problems for which there are no known solutions. Researchers must decide: (1) when to apply or extend theory, (2) when to apply data analytic techniques to extract and explore data patterns, and (3) when to do both. West (2013) describes such problems as "complex" because they have many different parts and interact in many different ways. For example: "What should we do about uncertainty in the financial markets? How can we predict energy supply and demand? How will climate change play out? How do we cope with rapid urbanization?" (West, 2013, p.1). Such problems are sufficiently complex that neither approach can work satisfactorily in isolation.

4.3.1 Resolution by Problem Refinement

To solve a complex problem, researchers need to refine it in an attempt to make it more tractable. It might be possible to simplify, or work with specific components of, the problem. For example, researchers can manage complexity (Kaul, Storey, & Woo, 2017) using techniques such as decomposing the problem, visualizing the different components of the problem, changing the parameters considered, articulating the scope, or eliminating constraints. If existing techniques used in data analytics cannot perform the analyses needed, then we need to extend or refine them. For example, the advent of social media required new text mining algorithms and techniques to process (and abstract) large amounts of unstructured data (e.g. Chua, Li, Kaul, & Storey, 2016; Ram, Zhang, Williams, & Pengetnze, 2015). Analogously, if a theory does not hold for a given problem, then it becomes necessary to adapt or modify the theory to deal with the exception.

4.3.2 Example

Suppose one wants to understand investor sentiment regarding the health of the stock market and the effect of such sentiment on actual investments. This is a complex problem that cannot be solved by traditional finance theories because they are limited to specific terms that appear in outlets such as media coverage, investment analysts' reports, and earnings announcements. It also cannot be solved by data/text mining or sentiment analysis because these techniques cannot explain, for example, why a certain trading strategy works and why a set of events affects the stock market (Gu, Storey, & Woo, 2015).

To address the complexity of this problem, one needs to understand the relevant theories and data analytic techniques that are available and how they might be adapted. One possible approach is to visualize all of the possibilities using a diagram or table for analysis. For example, the business intelligence model (Horkoff et al., 2014) was developed as a general conceptual model for representing business needs, which can also be applied to capture data analytic capabilities.

4.3.3 Role of Information Systems Research

Information systems researchers can propose conceptual models to analyze the matching of theory and techniques with a problem to be solved. In the example above, Gu et al. (2015) applied the business intelligence model to match financial theories to data analytics techniques. However, this conceptual model could not capture different scenarios (competing alternatives) in the same diagram. To resolve this problem, the business intelligence model might be extended to allow for multiple scenarios, highlighting an insufficiency in matching financial theories with data analytics techniques. Thus, the resolution of this issue requires refining a new modeling method (the business intelligence model) for representing important concepts in the financial domain (e.g., long-term versus short-term effects on the stock market). The role of IS research is to develop, test and refine conceptual and other models, playing a *responsive* role driven by a complex problem. The roles that IS researchers can play are further elaborated using examples in Table 4.

Table 4. Roles of IS Researchers Responding to Challenge 3: Addressing Problems That Are Unsolvable from A Single Perspective

Research direction	What IS researchers can contribute
Develop approaches for bridging specific practical problems and general theoretical problems	Apply, extend and develop research methodologies, such as action research, that help bridge the gap between data-driven and theory-driven perspectives (Sein, Henfridsson, Purao, Rossi, & Lindgren, 2011; Davison, Martinsons, & Ou, 2012)
Develop strategies for managing the complexity that results from bridging data-driven and theory-driven perspectives	Apply, extend, and develop techniques, such as scoping and decomposition, that simplify complexity (Burton-Jones & Meso, 2006; Kaul, Storey, & Woo, 2017a)
Develop modeling approaches that allow representing, reasoning with, and combining possible solutions to problems	Apply, extend, and develop multi-perspective conceptual modeling approaches (Paja, Maté, Woo, & Mylopoulos, 2016)

4.4 Challenge 4: Sharing Data and Models across Research Teams and Projects

Data sharing in any research environment means using data created by others. The challenge is in providing contextual information and resolving data heterogeneity (variety). Prior to the era of big data, researchers working with domain theories identified the data they required for theory-testing and created or acquired the necessary instruments to collect it before carrying out the actual data collection. Groves (2011) refers to this type of data as "designed data." However, data designed for one research study is not easily shared for use in another (Jarvenpaa & Staples, 2000). This issue is exacerbated in the big data era, where "organic data" (Groves, 2011) is typically generated without identified objectives for analysis. For example, social media data is generated independently of research objectives, but can be repurposed to address specific research questions (e.g., Ram et al. 2015). Historically, sharing designed (or organic) data has not been a normal part of research activity.

Several initiatives have been proposed to create the infrastructures needed to share data. For example, the Cancer Biomedical Informatics Grid (caBIG) and National Cancer Informatics Program (NCIP) provide access to cancer-related data. This is enabled by metadata models and algorithms for sharing and collaborating (https://cbiit.nci.nih.gov/ncip/about-ncip/mission). However, there is no similar infrastructure for sharing organic data or combining designed and organic data.

4.4.1 Resolution by Preparing Data for Future Use

The challenge of sharing data across research teams and projects can be resolved by capturing and representing data in meaningful ways for ease of understanding and interpretation, thus facilitating its availability for future projects. The use of appropriate data is a key issue in both data analytics and domain theory development. In data analytics, the content and quality of data are crucial for selecting, integrating, and manipulating data from heterogeneous sources and with different formats. For theory development, the main activities involve specifying the meaning (semantics) of theoretical constructs and their relationships to each other.

4.4.2 Example

Wang, Mai, & Chiang (2014) attempted to make manufacturer-provided content of tablet computer data meaningful. To do so, they organized the user-generated content into four classes (the market dynamics of products, product characteristic information, consumer-generated product reviews, and reviewer information), with detailed descriptions of the attributes for each class. This enables others to conduct research on relevant topics (e.g., product reviews, pricing, competition, new product development, and text analytics), without being concerned with cleaning, conceptualizing, and integrating the data.

4.4.3 Role of Information Systems Research

IS researchers can apply and extend their work in data quality and data integration of heterogeneous data to large datasets to facilitate data sharing. Data quality greatly affects the feasibility of future use of data. Research has focused on understanding how data

quality can be ensured through controls on data entry and interpretations of whether the data are suitable for specific purposes (Wang et al., 1995; Wang and Strong, 1996). Data analytics increasingly relies on user-generated content instead of, or in combination with, traditional organizational data. User-generated content often can be created by anyone, in any format, with possible inconsistencies, uncertain quality, and different or conflicting meanings. This makes the focus on data quality and interpretation critical. When combining data from independent sources, semantic data integration techniques need to be applied (e.g., Goh, Bressan, Madnick, & Siegel, 1999) in ways that ensure data consistency and veracity. Collectively, these IS research initiatives contribute greatly to the selection, evaluation, and integration of data sources prior to applying data analytic techniques, thus ensuring that the data are useful for continued use. IS researchers can facilitate this kind of data sharing. In this manner, IS researchers can play a *proactive* role in anticipating how big data can be made usable for future, possibly unspecified, needs. The roles that IS researchers can play are further elaborated using examples in Table 5.

Table 5. Roles of IS Researchers Responding to Challenge 4: Share Data, Models, and Workflows Across Research Teams and Projects

Research direction	What IS researchers can contribute
Support the sharing of data, models, and workflows across different domains and perspectives	Apply, extend, and develop "open science" principles to support the sharing of data and models via infrastructure and platforms (Fecher & Friesike, 2013)
Develop approaches for curating data sources, models, and workflows for future uses	Apply and extend instance-based representation (Parsons, 1996)
Develop and maintain mapping and derivations across data sources, analytic models, and workflows	Apply and extend approaches for managing data dictionaries and model repositories (Smirnov, Weidlich, Mendling, & Weske, 2012)

4.5 Summary of Roles for Information Systems Researchers

Information systems research can play several roles that facilitate interactions between data-driven and theory-driven research in addressing the challenges posed in Table 1. First, to facilitate the paths from data analytics to domain theory, and, inversely, from domain theory to data analytics, IS scholars can play an *intermediary* role in which they link work from both perspectives to gain mutual benefits. As intermediaries, IS researchers can be responsible for understanding the work done in each perspective and translating the values held by researchers in each—a role similar to the liaison role they have advocated over the years for systems analysts. Second, IS researchers can play a *responsive* role in which they adapt data analytics techniques and domain theory to solve a problem. In this role, IS scholars can respond to the demands of a problem situation by identifying results generated by researchers working within each perspective that may be appropriate for the problem at hand. Finally, IS scholars can play a *proactive* role in which they prepare data for future problems, needs, or changes. As part of this role, IS researchers will need to anticipate concerns that go beyond a single research project to generate approaches and infrastructures that build capacity for, and facilitate work across, multiple research settings and projects (e.g., to address problems such as data sharing). Table 6 summarizes the resolutions to the four challenges, specifying the role of IS in each.

Table 6. Resolution of Challenges at The Intersection of Data-Driven and Theory-Driven Research

Challenge	Resolution	Role of IS research
Reconciling competing approaches to creating or defining domain theories using big data	Modeling to abstract and generalize patterns from data analytics to support developing and refining theory	Intermediary role • Provide modeling approaches (conceptual modeling, ontology development) to support explanation, interpretation, and generalization of high-level concepts derived from patterns
Selecting data and analytic techniques to conduct theory testing	Gathering and assessing authentic data requirements from theoretical constructs	Intermediary role • Apply requirements engineering to identify, select, and prepare relevant data to move from domain theory to data analytics
Solving problems that are unsolvable from a single perspective	Problem refinement to manage complexity	Responsive role • Refine complex problems to make them more tractable by applying systems analysis approaches and conceptual modeling (e.g., defining scope, managing complexity)
Sharing data across research teams and projects	Preparing data for future use, including data integration (combining design and organic data) and data quality assessment	Proactive role • Organize and prepare data for future use by applying and extending work on data quality and data integration of heterogeneous data

5 Conclusion

This paper has developed a framework for information systems research to facilitate interaction between the data-driven and theory-driven research. The framework represents activities related to the intrinsic differences[1] in how data-driven and theory-driven research are conducted to identify four specific challenges: reconciling the two perspectives (data-driven and theory-driven research), selecting data and techniques for theory testing, solving unsolvable problems, and sharing data. To respond to these challenges, information systems researchers can play intermediary roles (for the first two perspectives), as well as responsive and proactive roles. The challenges present opportunities for information systems researchers to apply or extend knowledge of systems analysis (including requirements engineering and managing complexity), conceptual modeling (abstractions, generalizations), ontology development, data quality, and data integration. Pursuing these opportunities will establish important research interactions between data-driven and theory-driven perspectives, thus contributing to research success in the big data era.

[1] Such differences across communities of research and practice have been identified and addressed by scholars starting with the work of Snow (1959) and Breiman (2001).

Acknowledgments

The authors wish to thank the editor in chief and the three reviewers whose constructive feedback contributed significantly to this paper. This research was supported by the Big Data Innovation Network, Bentley University; J. Mack Robinson College of Business, Georgia State University; the Natural Sciences and Engineering Research Council of Canada; and the Deutsche Forschungsgemeinschaft (DFG).

References

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems*, *17*(2), i-xxxii.

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

Andersen, H. & Hepburn, B. (2016). Scientific method. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/entries/scientific-method/

Avison, D., & Fitzgerald, G. (2003). *Information systems development: Methodologies, techniques and tools*. New York, NY: McGraw Hill.

Baesens, B., Bapna, R., Marsden, J.R., Vanthienen, J., & Zhao, J. L. (2016). Transformational issues of big data and analytics in networked business. *MIS Quarterly*, *40*(4), 807-818.

Baldoni, M., Baroglio, C., Patti, V., & Rena, P. (2012). From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, *6*(1), 41–54.

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, *3*(3), 274–279.

Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, *323*(5919), 1297–1298.

Bentley, R.A., O'Brien M. J., & Brock W. A. (2014). mapping collective behavior in the big-data era. *Behavioral and Brain Sciences, 37*(1), 63–76.

Blouin, J., Bard, C., Teasdale, N., Paillard, J., Fleury, M., Forget, R., & Lamarre, Y. (1993). Reference systems for coding spatial information in normal subjects and a deafferented patient. *Experimental Brain Research*, *93*(2), 324–331.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 99–231.

Chan, J., Ghose, A., & Seamans, R. (2016). The internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly*, *40*(2), 381–403.

Chiang, R. H., Barron, T. M., & Storey, V. C. (1994). Reverse engineering of relational databases: Extraction of an EER model from a relational database. *Data & Knowledge Engineering*, *12*(2), 107–142.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications, 83*, 187–205.

Chua, C., Li, X. D., Kaul, M., & Storey, V. C. (2016, May). *Mining social media data from sparse text: An application to diplomacy.* Paper presented at the 11th International Conference on Design Science Research in Information Systems and Technology, St. John, Canada.

Cruz, J. A., & Wishart, D.S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics 2*, 59-78.

Davenport, T. H., Barth, P., & Bean, R. (2013). How "big data" is different. *MIT Sloan Management Review*, *54*(1), 22-24.

Davison, R. M., Martinsons, M. G., & Ou, C. X. (2012). The roles of theory in canonical action research. *MIS Quarterly, 36*(3), 763–786.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64–73.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, *4*(1), eaao5580, 5pp.

Economist. (2017, October). The latest AI can work things out without being taught. *Economist.* Retrieved from https://www.economist.com/science-and-technology/2017/10/21/the-latest-ai-can-work-things-out-without-being-taught

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532–550.

Evermann, J., & Hallimi, H. (2008). Associations and mutual properties: An experimental assessment. *Proceedings of the Americas Conference on Information Systems*. AIS

Fecher B., & Friesike S. (2014). Open science: One term, five schools of thought. In: S. Bartling & S. Friesike (Eds.), *Opening Science*. New York, NY: Springer.

Finucane, M. M., Paciorek, C. J., Danaei, G., & Ezzati, M. (2014). Bayesian estimation of population-level trends in measures of health status. *Statistical Science*, *29*(1), 18–25.

Goes, P. (2014). Editor's comments: Big data and IS research. *MIS Quarterly, 38*(3). iii–viii.

Goh, C. H., Bressan, S., Madnick, S., & Siegel, M. (1999). Context interchange: New features and formalisms for the intelligent integration of information. *ACM Transactions on Information Systems*, *17*(3), 270–293.

Golledge, R. G. (1999). Human wayfinding and cognitive maps. In R.G. Golledge (Ed.), *Wayfinding behavior: Cognitive mapping and other spatial processes* (pp. 5–45). Baltimore, MD: Johns Hopkins University Press.

Greenwood, B., & Agarwal, R. (2016). Matching Platforms and HIV Incidence: An Empirical Investigation of Race, Gender, and Socioeconomic Status. *Management Science*, *62*(8), 2281–2303

Groves, R. (2011). Three eras of survey research: Designed versus organic data. *Public Opinion Quarterly*, *75*(5), 861–871.

Gu, Y., Storey, V. C., & Woo, C. C. (2015). Conceptual modeling for financial investment with text mining. *Proceedings of the 34th International Conference on Conceptual Modeling* (pp. 528–535). Springer.

Günther, W. A., Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems 26*(3)*,* 191–209.

Gupta, A., Deokar, A., Iyer, L., Sharda, R., & Schrader, D. (2018). Big data & analytics for societal impact: Recent research and trends. *Information Systems Frontiers*, *20*(2), 185–194.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., . . . & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162.

Hastie, T., R., Tibshirani, & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York, NY: Springer.

Horkoff, J., Barone, D., Jiang, L., Yu, E., Amyot, D., Borgida, A., & Mylopoulos, J. (2014). Strategic business modeling: Representation and reasoning. *Software & Systems Modeling 13*(3), 1015–1041.

Jagadish, H. V. (2015). Big data and science: Myths and reality. *Big Data Research*, *2*(2), 49–52.

Jarvenpaa, S. L., & Staples, D. S. (2000). The use of collaborative electronic media for information sharing: An exploratory study of determinants. *The Journal of Strategic Information Systems*, *9*(2), 129–154.

Kaul, M., Storey, V. C., & Woo, C. (2017). A Framework for managing complexity in information systems. *Journal of Database Management*, *28*(1), 31–42.

Kaul, M., Storey, V. C., & Woo, C. (2017a). Domain design principles for managing complexity in conceptual modeling. *Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology*. Design Science.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 1–12.

Kotonya, G., & Sommerville, I. (1998). *Requirements engineering: Processes and techniques*. Hoboken, NJ: Wiley.

Lazer, D., & Kennedy, R. (2015, October). What we can learn from the epic failure of Google Flu Trends. *Wired*. Retrieved from https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science 343*(6176), 1203–1205.

Lukyanenko, R., Samuel, B., & Parsons J. (2018). Artifact Sampling: Using Multiple Information Technology Artifacts to Increase Research Rigor. *Proceedings of the 51st Hawaii International Conference on Systems Sciences*. AIS.

Maass, W., Parsons, J., Purao, S., Rosales, A., Storey, V. C., & Woo, C. C. (2017). Big data and theory. In L Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1–5). New York, NY: Springer Nature.

Madnick, S., & Zhu, H. (2006). Improving data quality through effective use of data semantics. *Data & Knowledge Engineering*, *59*(2), 460–475.

Markus, L., & Topi, H. (2015). Big data, big decisions for science, society, and business: Report on a research agenda setting workshop Retrieved from https://www.bentley.edu/files/2015/10/08/BigDataWorkshopFinalReport.pdf

Mathiassen, L., & Purao, S. (2002). Educating reflective systems developers. *Information Systems Journal*, *12*(2), 81–102.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

NCD Risk Factor Collaboration. (2016). Trends in adult body-mass index in 200 countries from 1975 to 2014: A pooled analysis of 1698 population-based measurement studies with 19· 2 million participants. *The Lancet*, *387*(10026), 1377–1396.

Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*, *10*, 4915–4931.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943–951.

Paja, E., Maté, A., Woo, C., & Mylopoulos, J. (2016). Can goal reasoning techniques be used for strategic decision-making? *Proceedings of the 35th International Conference on Conceptual Modeling* (pp. 530–543). Springer.

Pankratius, V., & C. Mattmann (2014). Computing in astronomy: To see the unseen. *Computer 47*(9), 23–25.

Parsons, J. (1996). An information model based on classification theory. *Management Science*, *42*(10), 1437–1453.

Parsons, J., & Wand, Y. (2013). Extending principles of classification from information modeling to other disciplines. *Journal of the Association for Information Systems,* 14(5), 245–273.

Rai, A. (2016). Synergies Between Big data and theory. *MIS Quarterly*, *40*(2), iii–ix.

Ram, S., Zhang, W., Williams, M., & Pengetnze, Y. (2015). Predicting asthma-related emergency department visits using big data. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1216–1223.

Sein, M.K., Henfridsson, O., Purao, S., Rossi, M.. & Lindgren, R., (2011). Action design research. *MIS Quarterly*, *35*(1), 37–56.

Sellars, S., Nguyen, P. Chu, W. Gao, X., Hsu, K., & Sorooshian S. (2013). Computational earth science: Big data transformed into insight. *Eos, Transactions, American Geophysical Union*, *94*(32), 277–288.

Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive Psychology, 43*(4), 274–310.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.

Siegfried, T. (2013, December). Rise of big data underscores need for theory. *Science News*. https://www.sciencenews.org/blog/context/rise-big-data-underscores-need-theory/

Silver, David, et al. (2017). Mastering the game of go without human knowledge. *Nature 550*(7676), 354–359.

Smirnov, S., Weidlich, M., Mendling, J., & Weske, M. Action patterns in business process model repositories. *Computers in Industry 63*(2), 98–111.

Snow, C. P. (1959). The two cultures and the scientific revolution. Cambridge, U.K.: Cambridge University Press.

Su, K., Huang, H., Wu, X., & Zhang, S., 2006. A logical framework for identifying quality knowledge from different data sources. *Decision Support Systems*, *42*(3), 1673–1683.

Sugumaran, V., & Storey, V. C. (2002). Ontologies for conceptual modeling: Their creation, use, and management. *Data & Knowledge Engineering*, *42*(3), 251–271.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: Massachusetts Institute of Technology Press.

Unger, D. G., & Wandersman A. (1985). The importance of neighbors: The social, cognitive, and affective components of neighboring. *American Journal of Community Psychology, 13*(2), 139–169.

Vidovic, M. M-C., et al. (2015). Opening the black box: Revealing interpretable sequence motifs in kernel-based learning algorithms. *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Part II*. Springer.

Wand, Y., Monarchi, D. E., Parsons, J., & Woo, C. C. (1995). Theoretical foundations for conceptual modelling in information systems development. *Decision Support Systems*, *15*(4), 285–304.

Wand, Y., Storey, V. C., & Weber, R. (1999). An ontological analysis of the relationship construct in conceptual modeling. *ACM Transactions on Database Systems*, *24*(4), 494–528.

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering, 7*(4), 623–640.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 5–33.

Wang, X., Mai, F., Chiang, R. H (2014). Market dynamics and user-generated content about tablet computers. *Marketing Science,* 33(3), 449–458.

Weber, R. (2003). Editor's comment: Still desperately seeking the IT artifact, *MIS Quarterly*, *27*(2), iii–xi.

West, G. (2013). Big data needs a big theory to go with it. *Scientific American*, *308*(5). Retrieved from http://www.scientificamerican.com/article/big-data-needs-big-theory/.

Woo, C. (2011): The role of conceptual modeling in managing and changing the business. *Proceedings of the 30th International Conference on Conceptual Modeling*. Springer

Woods, J., & Rosales, A. (2010). Virtuous Distortion. In L. Magnani, W. Carnielli, & C. Pizzi (Eds.), *Model-based reasoning in science and technology* (pp. 3–30). Berlin: Springer.

Xiao, B., & Benbasat, I. (2015). Designing warning messages for detecting biased online product recommendations: An empirical investigation. *Information Systems Research*, *26*(4), 793–811.

About the Authors

Wolfgang Maass is a professor of business informatics and a professor of computer science (joint appointment) at Saarland University; scientific director at the German Research Center for Artificial Intelligence (DFKI); and an adjunct professor at Stony Brook University, School of Medicine. He studied computer science at RWTH Aachen and Saarland University. His PhD in computer science at Saarland University was funded by the German National Science Foundation (DFG); he investigated incremental natural language route descriptions. He was a postdoc researcher at the University of St. Gallen, where he received his *Habilitation* from the Department of Management. Previously he was a professor of media informatics at Furtwangen University and a guest professor in the Department of Bioinformatics and Computational Biology at MD Anderson Cancer Center, University of Texas, and in the Department for Biomedical Informatics at Stony Brook University Health Sciences Center School of Medicine. His research investigates the relationship between conceptual modeling and AI/ML, AI in distributed environments (EdgeAI), and the design of smart services. Additionally he has a strong focus on knowledge transfer to industry through funded projects and spin-offs.

Jeffrey Parsons is University Research Professor and professor of information systems in the Faculty of Business Administration at Memorial University of Newfoundland. His research interests include conceptual modeling, crowdsourcing, information quality, data integration, and recommender systems. His work on these topics has appeared in top journals in information systems (e.g., *Information Systems Research, Journal of the Association for Information Systems, MIS Quarterly*), management (e.g., *Management Science*), computer science (e.g., *ACM Transactions on Database Systems, IEEE Transactions on Knowledge and Data Engineering*), and biology (e.g., *Nature, Conservation Biology*). He is a senior editor for *MIS Quarterly*, a former senior editor for the *Journal of the Association for Information System*, and he has served as program co-chair for a number of major information systems conferences, including AMCIS, WITS, ER, and DESRIST. He has received numerous awards, including the INFORMS ISS Design Science Award and the designation of ER Fellow.

Sandeep Purao is Trustee Professor in the Information and Process Management Group at Bentley University. Prior to joining Bentley, he was a professor in the College of Information Sciences and Technology at Penn State University and was on the faculty of the Business School at Georgia State University. His research focuses on the design and evolution of complex techno-organizational systems; and the sciences of design. His work has been published in journals such as *MIS Quarterly*, *Communications of the ACM*, various *IEEE Transactions*, *ACM Computing Surveys*, and *Information Systems Research*; and has been presented at conferences such as the International Conference on Information Systems, IEEE Service-oriented Computing Conference, and International Conference on Conceptual Modeling. He serves or has served in editorial capacities for *Information Systems Research*, *MIS Quarterly*, *Journal of the Association for Information Systems*, *European Journal of Information Systems*, and *Scandinavian Journal of Information Systems*. He has held leadership roles for conferences such as IEEE Services, ER, WITS and DESRIST. His research has been funded by the National Science Foundation, private foundations, and industry consortia. He holds a PhD in management information systems from the University of Wisconsin-Milwaukee. He is a member of AIS, ACM, and IEEE.

Veda C. Storey is the Tull Professor of Computer Information Systems and professor of computer science at the J. Mack Robinson College of Business, Georgia State University. Her research interests are in intelligent information systems, data management, conceptual modeling, and design science research. Dr. Storey is a member of the AIS College of Senior Scholars, an AIS Fellow, and an advisor to the Workshop on Information Technologies and Systems. She is also a member of the steering committee of the International Conference of Conceptual Modeling, where she has the honor of being an ER Fellow and a recipient of the Peter P. Chen Award. She received a Georgia State University Teaching Innovation Award for her work on experiential and interdisciplinary teaching. Dr. Storey received her PhD from the University of British Columbia and holds a degree in flute performance from the Royal Conservatory of Music, University of Toronto.

Carson Woo is Stanley Kwok Professor of Business, Sauder School of Business, University of British Columbia. His research interests include conceptual modeling, systems analysis and design, and requirements engineering. In particular, he is interested in using conceptual models to acquire relevant contextual information (e.g., business goals) and utilizing it to design new information systems, or aligning it to existing information systems design, so that changes can be more appropriate to business needs. At the University of British Columbia, he is a member of two research clusters: (1) Artificial Intelligence, and (2) Blockchain. Dr. Woo is editor of *Information Technology and Systems Abstracts Journal at the Social Science Research Network*, and serves or has served on several editorial boards, including *ACM Transactions on Management Information Systems, Business & Information Systems Engineering Journal, Information and Management,* and *Requirements Engineering*. He currently serves as the chair (2019–2020) of the Conceptual Modeling Conference steering committee and has served as the president of the Workshop on Information Technology and Systems (2004-2006) and chair of the ACM Special Interest Group on Office Information Systems (1991-1995).