Rizk and Elragal *J Big Data* (2020) 7:7
https://doi.org/10.1186/s40537-019-0280-6

Journal of Big Data

RESEARCH Open Access

Data science: developing theoretical contributions in information systems via text analytics

Aya Rizk* and Ahmed Elragal

*Correspondence:
aya.rizk@ltu.se
Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, 97187 Luleå, Sweden

Abstract

Scholars have been increasingly calling for innovative research in the organizational sciences in general, and the information systems (IS) field in specific, one that breaks from the dominance of gap-spotting and specific methodical confinements. Hence, pushing the boundaries of information systems is needed, and one way to do so is by relying more on data and less on a priori theory. Data, being considered one of the most important resources in research, and society at large, requires the application of scientific methods to extract valuable knowledge towards theoretical development. However, the nature of knowledge varies from a scientific discipline to another, and the views on data science (DS) studies are substantially diverse. These views vary from being seen as a new scientific (fourth) paradigm, to an extension of existing paradigms with new tools and methods, to a phenomenon or object of study. In this paper, we review these perspectives and expand on the view of data science as a methodology for scientific inquiry. Motivated by the IS discipline's history and accumulated knowledge in using DS methods for understanding organizational and societal phenomena, IS theory and theoretical contributions are given particular attention as the key outcome of adopting such methodology. Exemplar studies are analyzed to show how rigor can be achieved, and an illustrative example using text analytics to study digital innovation is provided to guide researchers.

Keywords: Data science, Theory, Contribution, Information systems, Text analytics, Methodology

Introduction

Just like in business and society, data in research is increasing in volume, velocity and variety, and requires new ways of extracting value from it. Data science (DS)—the systematic extraction of knowledge from data—has been attracting a lot of attention recently [1]. It is argued that data science is leading a new scientific paradigm [2, 3]. Its epistemological assumptions, challenges and opportunities have been discussed in various disciplines [4, 5]. However, there are also questions about whether it is really a new (fourth) paradigm of science or empiricism re-emerging [6], or simply an extension of existing paradigms with new tools and methods for scientific enquiry [7].

The views in the Information Systems (IS) are equally diverse. Data science is viewed as a paradigm [8, 9], a methodology [4], a method [10], or a phenomenon of study [11]. These approaches primarily discuss opportunities and challenges of adopting data science for scientific discovery. However, a key element in scientific discovery, that is theory and the nature of theoretical knowledge, is often ignored. In this paper, these different views of data science and how they are adopted in IS research are presented. Furthermore, we expand on viewing data science as a methodology for generating theoretical contributions in the IS discipline.

This research is motivated by a few elements. First, the abundance of data that captures social events and activities, being ever so close to phenomena they represent, requires us to discuss new approaches to theorizing [12]. Second, advancements in analytical capabilities and computational methods allow for richer understanding that enables such theory development endeavors [6, 13]. Third, one response to the dire need for innovative research is relying more on data and less on a priori theory [14, 15]. Fourth, the IS discipline is especially well suited to lead discussions on (organizational and social) theory development via data science [1].

To this end, we first review the nature of theory and theoretical contributions in IS, followed by a brief discussion on data science and its state in the field. Then we examine data science contributions in IS in the last 5 years and argue for DS as a research methodology. Next, we describe this methodology with guidelines towards building a variety of theoretical contributions from DS studies. Finally, we provide a practical example for better grounding before we conclude.

Background

Theory in information systems

Discussion on what theory is and what it is not is crucial towards the development of any discipline. In IS, it has been particularly difficult to describe the structure of IS theories since the discipline deals with phenomena arising at the intersection of the natural, social, and artificial (design) sciences [16, 17]. This key structural and ontological question has some answers though. Theory in IS has been described in terms of what it constitutes, what it represents and what it intends to achieve, each addressed in turn as follows.

In the simplest form, a theory is comprised of a set of statements. These statements are language-bound, capture specific concepts—including constructs, units, factors and variables, and make a claim or a proposition about relationships between those concepts [16, 18]. Optionally, these statements may be complemented by other means of representation, such as tables, diagrams, graphs, etc. Accordingly, the two key structural elements that constitute theoretical statements are concepts and propositions. In addition to structural elements, theories constitute assumptions about their underlying logic, temporal and contextual factors that specify their range of coverage, or boundaries of generalizability [19].

Concepts are ideas [20] that we are able to give names, and they are abstractions related to the objects or phenomena of study [18]. They are the basic units for making sense of the world [21]. Concepts are generally differentiated based on their level of abstraction—i.e. whether they can be observed or measured empirically, or not

[ibid]. Constructs are a specific type of concepts that are not observable themselves but should be fully defined in observable terms. Theoretical concepts are the most abstract and refer to concepts that cannot be measured or observed, and are typically theory-bound [18]. Variables, on the other hand, are operational and measurable configurations that are derived from concepts/constructs, and can assume two or more values [22]. Focusing on concepts enables any field to recognize its body of knowledge within a broader perspective and its value to its intended stakeholders. This allows scholars to address wider problems and advance alongside other areas of study. It also avoids the undesired path in which a theory is adopted so far from its origin and gets confused with other theories that come from a different system of thought and different set of assumptions [18].

The second key component of a theory that binds concepts together is propositions: a group of field-specific statements that define or relate concepts within that field [18]. The level of abstraction of propositions essentially depends on that of the constituent concepts. Gibbs [23] defines two types of propositions depending on their level of abstraction. First, postulates are propositions that contain observable concepts and can be tested. Second, axioms are propositions that contain abstract concepts and cannot be tested directly. Indeed, these are not mutually exclusive, and most often propositions contain both observable and unobservable concepts [16]. Now when propositions are connecting both theoretical and empirical languages, they are called epistemic statements. Hypotheses are special type of epistemic statements that make a claim about the data, including signs (i.e. positive or negative) and moderation. The scope of the theory and its generality should be defined using boundaries and modal qualifiers [16]. The set of statements constituting the theory should specify the extent of applicability of those statements using words such as "some, every, all, always" [16, p. 616], or the class of problems such knowledge intends to solve [24, 25]. The issue of generalizability has long been discussed, backed by extensive philosophical framework [26].

Theory has a function; that is to capture our complex world [18]. It is contemplative, abstract and is bounded by assumptions and constraints. It aims to "describe, explain, and enhance our understanding of the world and, in some cases, [to] provide predictions of what will happen in the future and [to] give a basis for intervention and action" [16, p. 616].

To contribute to a body of knowledge, theory needs to maintain coherence while progressively pushing the boundaries of the respective field. Grover and Lyytinen [15] argue that the theorizing practices currently dominating in the IS domain are limiting its potential and resulting in an incoherent discourse. Thus, Hassan et al. [27] recommend that we view theorizing as a discursive practice, where the key components of a theory are a product of traversing between foundational and generative theorizing practices.

Nevertheless, the way in which these discursive practices are conducted to organize theory, or theorize, essentially follows from the underlying goal of the theory in question. The primary goal of a theory describes what we intend to achieve by developing such a theory, and often follows from identified research problems and questions. Four primary goals of theoretical propositions in IS have been identified as analysis and description, explanation, prediction and prescription. Accordingly, Gregor [16] identifies five main types of theory—see Table 1 for a brief description.

Table 1 A taxonomy of theory types in information systems research [16, p. 620]

Theory type	Distinguishing attributes
I. Analysis	Says what is The theory does not extend beyond analysis and description. No causal relationships among phenomena are specified and no predictions are made
II. Explanation	Says what is, how, why, when and where The theory provides explanations but does not aim to predict with any precision. There are no testable propositions
III. Prediction	Says what is and what will be The theory provides predictions and has testable propositions but does not have well-developed justificatory causal explanations
IV. Explanation and prediction (EP)	Says what is, how, why, when, where and what will be Provides predictions and has both testable propositions and causal explanations
V. Design and action	Says how to do something The theory gives explicit prescriptions (e.g. methods, techniques, principles of form and function) for constructing an artifact

Table 2 Differences between Grand and Middle-range theories [18, p. 7]

Aspect	Grand theory	Middle-range theory
Boundary	Unbounded	Bounded by subject matter
Constitution	Axioms containing constructs and theoretical concepts	Propositions containing observables
Level of falsifiability	Low	High
Differentiated by	Philosophy	Specialization
Legitimacy	It is primarily a means of establishing legitimacy	Legitimacy is evidenced by scope, precision and investigative tools
Formation and growth	Fully formed from the mind of the theorist, and may grow as a result of discussion	Formed from a mass of basic observations, and grows by knowledge and experience of its scientists and researchers
Data and generalization	Does not require data, generalization is based on the paradox of induction	Requires data, but is abstract enough to provide generalization
Inception and systemic interactions	Starts from the outside with a total system and imposes on derived theories	Starts from the inside and possibly builds a unified system across domains

The role of theory in information systems has long been debated with regards to the field's identity, intellectual core, and the role of technological constructs in theoretical contributions and theorizing processes [18, 28]. With the dominant practice of borrowing grand theories from reference domains and introducing generic IT components, theoretical contributions in IS become mediocre, and building a coherent body of knowledge becomes very challenging [15].

Hassan and Lowry [18] suggest that targeting middle-range theories that are organic to the field of information systems would emancipate it from these challenges. Middle-range theories are defined as "logically interconnected sets of propositions that lie between concrete hypotheses and all-inclusive systematic efforts to explain all observed phenomena", or grand theories [18, p. 6]. Theories in the middle range, or substantive theory [16], differ considerably from grand or formal theories. Table 2 summarizes those differences. Perhaps two of the most important differences are the level of falsifiability (i.e. ability to test) and differentiation (i.e. bounded by philosophy or subject matter).

This paper deals with middle-range theories, those that are concrete enough to be empirically validated, and abstract enough to allow for generalization (in the broad sense). Before expanding on the contribution of data science to middle-range theories, data science is briefly introduced.

Data science in a nutshell

Data science is the study of systematic extraction of nonobvious and useful patterns and knowledge from data [29, 30] towards research advancement, organizational decision-making, and enabling a data-driven society [31]. However, this simple definition shows how much data science shares with science, scientific method, and business analytics. In this section, these terminologies are revisited, along with highlighting why data science is necessary as a distinct one. Unsurprisingly, data science shares much of the definition of science and scientific method, that is the "principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses" [32, p. 1]. Accordingly, it is what we define as *data*, the means of knowledge extraction and the end objective that set data science apart. For the purpose of this paper, data is viewed as a digital observation that serves as input to some research process.[1]

The means of knowledge extraction, on the other hand, would require an epistemological discussion. But before going into this discussion, clarification is necessary around what is meant by knowledge. In the data science domain and related domains (e.g. decision support systems), *knowledge* refers to insights and claims that can be made from analyzing factual data. However, scholars addressing meta-theoretical questions about what theory is and what it isn't, would not regard this reference to knowledge as outright theory [16]. Instead, these insights would need to go through what Hassan et al. [27] call a discursive practice of theorizing, towards producing *theoretical knowledge*, adding to the discipline's *body of knowledge*. While knowledge discovery may be the main goal of data science, building a coherent body of knowledge is one of the goals of every discipline.

On the epistemology of data science

Not all disciplines enjoy a unified view of accepted theories, philosophical, and methodological assumptions. Especially in the IS discipline, and more generally in the social sciences, a diverse range of philosophical assumptions are employed to make sense of socio-technical phenomena [9]. The IS discipline draws on traditions from the natural, social and artificial sciences [16], and on assumptions from the positivist, interpretive and critical epistemologies [33].

Positivist research strives for explanation and prediction, generalizing from sample to population. Interpretivist research seeks contextual interpretation and understanding of a phenomenon [34]. Critical research "focuses on the oppositions, conflicts and

[1] Research here is meant in an inclusive manner, where organization decision-making is also viewed as a research endeavor.

contradictions in contemporary society, and seeks to be emancipatory i.e. it should help to eliminate the causes of alienation and domination." [35, p. 5–6].

In a way, data science is viewed as a paradigm, one that comes with its own philosophical assumptions. New scientific paradigms emerge due to various reasons: (a) revolution in measurement [2], (b) advances in data generation, collection, management and analytical methods [3, 36], and/or (c) limitations in existing paradigms' ability in advancing knowledge discovery and theory development, or are not addressing key subject matter questions [37].

Whether scientific paradigm shifts are set in motion due to revolutions in measurement, advances in forms of data and associated analytical methods, or inability of existing paradigms to answer key questions, scholars agree that a fourth paradigm of science is emerging. This paradigm is associated with a data revolution, and the opportunities that come with it [1]. However, intensive debate has emerged on the nature of this paradigm, and why scholars need to be cautious not to fall in the same traps of empiricism [5]. Kitchin [9] summarizes the fallacies of empiricism that are paralleled in today's discussion around data science in the following claims [p. 133–137] (Table 3).

Accordingly, scholars advocate for data-intensive science [8], discovery science [38] or data-driven science [6], in which:

- It is suited to make sense of massive interconnected datasets, overcoming problems of small samples and scarce data
- Interdisciplinary research is fostered, since it is much less limited with a priori theoretical boundaries
- Abductive reasoning is encouraged
- Holistic models and theories about complex systems, rather than elements of it, are possible.

Every paradigm provides a different view on what valid theory is, and what a theory should do. This provides a lot of emphasis on methodological choices rather than on objects of study, at the risk of producing uninteresting and irrelevant knowledge [34].

Table 3 Fallacies of empiricism according to [9]

Claim	Counterargument
Big data (as a key force behind data science) can capture the full resolution of a given domain or phenomenon	No matter how exhaustive the data is, it is still a representation and a sample, and is bound by space and time in a continuously changing world
There is no need for a priori knowledge in the form of theory, models or hypotheses	Systems that capture and generate data are designed for very specific purposes, and analytical algorithms are designed through scientific reasoning (drawing on established theories)
Data can speak for themselves free from human bias	Patterns extracted from data require human interpretation and theorization to avoid "ecological fallacies" when taking action based on random correlations. Making sense of data is always framed through our knowledge and experiences
Meaning transcends context or domain-specific knowledge	Domain-specific expertise and wealth of knowledge is needed to assess and articulate problems and interpret results in order to avoid reductionism

While we visit methodological issues when reviewing data science contributions in IS, we first review common objectives of data science, which we believe bring it closer to objects of study and relevance. In the following section, these objectives are visited through looking at the commonalities, rather than the differences, among types of knowledge independent from their underlying paradigms.

Objectives of data science

Knowledge extraction from data is considered the high-level objective of data science. However, this objective is demonstrated differently in scholarly research as opposed to practice. Because theoretical knowledge is what sets scholarly research apart from practice, we distinguish between theoretical development and decision-making as two main objectives from data science efforts. Indeed, they are not mutually exclusive.

For theoretical development, the goal of the study (i.e. description, explanation, prediction or prescription) dictates the goals of data science. Some proponents of data science argue that prediction is its most powerful aspect [1, 29]. They also argue that it allows for a "state of knowing without understanding", where correlations and predictions are sufficient [39, p. 26]. On the other hand, common understanding of theory suggests that it has twofold objectives: explanation and prediction. Explanation lends itself to be the closest to human reasoning, and prediction allows both for theory testing and taking action. Even though most theories in IS are sought to achieve those two goals, they often achieve one at the expense of the other, leading to a paradox. Precision paradox occurs when a theory can predict but cannot explain, and power paradox occurs when a theory can explain but cannot predict with any degree of accuracy [16].

More practice-oriented data science efforts lean towards predictive and prescriptive knowledge extraction. Dhar [29] argues that utility of the extracted knowledge is paramount, that it is actionable for decision making. In this context, the ability of this new knowledge to predict the future supersedes its ability to explain the past events.

A rich and progressing body of knowledge, however, requires a diversity of goals, theories and approaches. Next, we review data science contributions in the IS discipline to show other views of data science and the potential to achieve such a diversity.

The state of data science in IS

Big data[2] and data science research are paramount to the advancement of IS research, and IS scholarship has the essential history and interdisciplinary position to take a leading role in data science research [40–42]. Rai [41] provides an excellent editorial on the synergies between big data and theory, where he elaborates four key areas where synergies between big data and theory can be achieved through research designs and methods as follows:

1. Theory informing the selection of constructs and concepts, their boundaries, and relationships among them.
2. Big data enhancing precision in theory testing.

[2] Big data is used to refer to the phenomenon of increasing volume, variety and velocity of data.

3. Diversity of datasets and analytical methods providing contextual elaboration of theories.
4. Generating theory for emergent phenomena.

Grover and Lyytinen [15] propose two approaches to overcome the limitations of the current practice of theorizing. The first approach is to focus on inductive data-driven research to uncover interesting and relevant patterns and concepts from data without heavy reliance on preexisting theory. The second approach lies in the other end of the empirical-theoretical continuum suggesting "blue ocean" theory development that does not necessarily base the research on data. Data science facilitates the former approach by offering the methods, tools and techniques to reveal regularities in data. Yet, theory-light approach here does not mean completely eliminating theory, for interpretation of such regularities does not happen in vacuum. And while data science may present the potential to overcome challenges with state-of-the-art theorizing, it requires significant attention to develop principles that would guide researchers on how to rigorously develop theories that are both relevant and interesting.

Scholars have already started engaging in such discussions. For instance, Müller et al. [7] provide guidelines for conducting rigorous studies in the field of information systems using big data analytics. Similarly, Elragal and Klischewski [4] identify the epistemological pitfalls associated with the process of conducting predictive data science studies and provide strategies to avoid those pitfalls along the research process. Thus far, these discussions are met with resistance due to, in part: (a) the difficulty of comprehending the results of most analytical techniques, (b) the scarcity of guidelines for such studies that enables researchers ensure rigor and validity (and reviewers assess it), and (c) the lack of guidelines enabling researchers to develop theoretical knowledge from data science results.

Debortoli et al. [43] offer a tutorial for researchers aiming to conduct such studies that would help them overcome the first two concerns. On the other hand, Berente et al. [13] propose a general approach addressing the latter concern, by drawing on elements of the manual grounded theory methods and automated computational theory discovery. The familiarity and acceptance of those two methods sets the ground for progressive discussions around data science, or computationally intensive, inductive theory development.

In addition to that, there is an increasing number of publications in IS that are considered data science studies. These studies unearth the presence of different views on data science, most of all empirical ones. To explore these views, a review is conducted as described in the following section.

Research method

A brief review of the IS literature from two journals is conducted to examine the scope of data science studies published between January 2015 and June 2019. The two selected journals are: (1) MIS Quarterly (MISQ), as one of the basket of eight[3] journals publishing general IS studies with focus on theoretical contributions, and (2) Journal of Big Data (JBD), as a specialized journal focused on data science issues both on theoretical and

[3] https://aisnet.org/page/SeniorScholarBasket.

practical levels. Due to that difference between the journals, the search strategies were slightly different.

For MISQ, the search included research articles with the terms "data science" or "analytics" in the abstract. Editorials, research notes, and commentaries were filtered out. As of July 10th, 2019, this search returned 20 papers. For JBD, due to the specialized nature of the journal, the search was extended to the keywords appearing anywhere in the text with 245 hits as of August 30th. A random sample of 25% (61 papers) was selected. Duplicates, conceptual, short papers, surveys and errata were excluded. Furthermore, we excluded studies addressing hardware and computing issues (e.g. [44]) and those including big data analytics as a construct in a theory testing framework (e.g. [45]), with a remainder of 13 and 31 papers from MISQ and JBD, respectively. The final set of papers included 44 papers from both journals.

Further examination of the papers aimed at understanding the nature of their contributions and the goals for such new knowledge. Each paper's abstract was read and both dimensions were extracted. Gregor's [16] theory goals discussed earlier are used to guide us classify the study's theoretical goals. Contributions were classified based on their tangible and utilitarian nature; that is whether the study produces theoretical contribution, designs an IT artifact, or a combination thereof. The aforementioned definition and constituents of theory are used to label if a study's contribution qualifies for theory. In case (one of) the study's contribution takes the form of a designed solution to a class of problems (e.g. system, algorithm, process, etc.), its contribution is (also) classified as an artifact [24, 25, 46]. In the discussion, we complement our findings by analyzing three studies adopting DS for theory development (see [13]).

Results: data science contributions to IS research

In addition to the paradigm perspective, the review reveals that data science is viewed as a methodology, a research method, and/or an object of study that constitutes (part of) the phenomenon of interest. We briefly describe each stream, the included articles, their main contributions and goal. Table 4 below summarizes the study type in each journal respectively.

Data science as a methodology (M)

In these studies, research questions and/or problems are addressed holistically via data science principles and procedures. Problems are either interdisciplinary (e.g. online branding or biomedicine) or from reference disciplines such as financial management, psychology or healthcare. They start with data, with little to no influence from theory,

Table 4 Classification of analytics and data science studies: 2015–2019

Scope of DS	MISQ	JBD	Total
Methodology	4	5	9
Research method	1	5	6
Object of study	2	14	16
Method and object	6	7	13
Total	13	31	44

and follow an inductive or abductive reasoning to knowledge generation. The rigor of these studies and the evaluation of the extracted knowledge stems from the analytical techniques chosen, and interpretation of the results takes place in reference to the respective discipline and body of knowledge. Examples from this stream of research (presented in Table 5 below) include identifying and predicting stressful behavior from mobile social data, designing a new method for forecasting commodity prices in the short-term, or explaining how retweets shape online personas.

The variety of techniques used in this stream highlight the methodological diversity that comes with using data science, as well as the diversity of the yielded theoretical contributions. Techniques used here include Bayesian approaches, decision trees, association rule mining and topic modeling, enabling scholars to achieve the wide range of theory goals. In addition, some studies use different techniques to address different subproblems or answer different sub-questions. Overall, this stream is balanced in terms of developing theoretical contributions and providing practical recommendations on designing data science models.

Data science as a research method (RM)

In this stream, the scope of data science is limited to a specific task in the research study, while the whole study adopts a different methodology. For instance, in [47–50] the authors prove that using specific techniques, combining specific datasets or features significantly improves predictive accuracy in their respective application domains. Note

Table 5 Sample of studies using data science as a methodology

Title of M-type studies	Contribution	Goal
Cerchiello, P., & Giudici, P. (2016). Big data analysis for financial risk management. *Journal of Big Data*, *3*(1), 18	Theory	Prediction
Hurtado, J. L., Agarwal, A., & Zhu, X. (2016). Topic discovery and future trend forecasting for texts. *Journal of Big Data*, *3*(1), 7	Theory	Prediction
Padmaja, B., Prasad, V. V. R., & Sunitha, K. V. N. (2016). TreeNet analysis of human stress behavior using socio-mobile data. *Journal of Big Data*, *3*(1), 24	Theory	Explanation and prediction
van Altena, A. J., Moerland, P. D., Zwinderman, A. H., & Olabarriaga, S. D. (2016). Understanding big data themes from scientific biomedical literature through topic modeling. *Journal of Big Data*, *3*(1), 23	Theory	Analysis
Wu, H., Wu, H., Zhu, M., Chen, W., & Chen, W. (2017). A new method of large-scale short-term forecasting of agricultural commodity prices: Illustrated by the case of agricultural markets in Beijing. *Journal of Big Data*, *4*(1), 1	Theory and artifact	Prediction and design
Geva, H., Oestreicher-Singer, G., & Saar-Tsechansky, M. (2019). Using Retweets When Shaping Our Online Persona: Topic Modeling Approach. *MIS Quarterly*, *43*(2)	Theory	Explanation
Gong, J., Abhishek, V., & Li, B. (2018). Examining the Impact of Keyword Ambiguity on Search Advertising Performance: A Topic Model Approach. *MIS Quarterly, 42*(3), 805–829	Theory and artifact	Explanation and design
Yahav, I., Shmueli, G., & Mani, D. (2016). A tree-based approach for addressing self-selection in impact studies with big data. *MIS Quarterly*, *40*(4), 819–848	Theory and artifact	Design
Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence. *MIS quarterly*, *40*(4), 1035–1056	Theory and artifact	Design

Table 6 Sample of studies using data science as a research method

Title of RM-type studies	Contribution	Goal
Agarwal, A., Baechle, C., Behara, R. S., & Rao, V. (2016). Multimethod approach to wellness predictive modeling. *Journal of Big Data*, *3*(1), 15	Theory and artifact	Prediction and design
Asri, H., Mousannif, H., & Al Moatassime, H. (2019). Reality mining and predictive analytics for building smart applications. *Journal of Big Data*, *6*(1), 66	Theory and artifact	Prediction and design
Goswami, K., Park, Y., & Song, C. (2017). Impact of reviewer social interaction on online consumer review fraud detection. *Journal of Big Data*, *4*(1), 15	Theory	Explanation and prediction
Mavragani, A., & Ochoa, G. (2018). Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. *Journal of Big Data*, *5*(1), 30	Theory	Prediction
Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*, *5*(1), 3	Theory	Prediction
Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forum and search data for sales prediction of high-involvement projects. *MIS Quarterly*, *41*(1), 65–82	Theory	Prediction

that the overall research design takes on an experimental approach. On the other hand, [51, 52] adopt a systems development approach to their studies, where predictive modeling is a sub-task. They provide prototypes as their main outcomes, thereby providing design recommendations along with their predictive contributions. This stream is the least represented in our sample, with just six studies (see Table 6).

Data science as an object of study (O)

In this stream, data science is regarded as an object of study rather than a method or methodology. Accordingly, studies in this stream are predominantly applied, technical and follow principles from other methodologies such as experimental design, design science research, or statistical hypothesis testing. With one exception following the latter methodology, studies in this stream address problems with existing data science techniques and algorithms or propose new ones through design and evaluation. Since the outcome of these studies take on the form of artifacts (i.e. framework, method, model, system architecture…etc.), their contributions take on two flavors: (1) design principles and prescriptive statements of how to build a similar one, and/or (2) action in case the artifact was tested through an intervention and evaluation in real-life context (see sample in Table 7).

Data science as a research method and object of study (RMO)

Studies in this class are similar to those in the M and RM classes in the sense that they set out to solve problems or answer questions using data science methods (among others) while applying rigor measures from other approaches. However, they also contribute to the data science domain through an artifact they develop throughout the research study (similar to studies in O). Accordingly, the role of theory varies considerably among these studies: from directly influencing feature selection in recommendation systems [53] to using it as an analytic lens throughout design and evaluation [54] (see Table 8).

Table 7 Sample of studies of data science as an object of study

Title of O-type studies	Contribution	Goal
Chandak, M. B. (2016). Role of big-data in classification and novel class detection in data streams. *Journal of Big Data*, *3*(1), 5	Artifact	Design
Chopade, P., & Zhan, J. (2015). Structural and functional analytics for community detection in large-scale complex networks. *Journal of Big Data*, *2*(1), 11	Artifact	Design
Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, *2*(1), 5	Artifact	Design
Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data*, *6*(1), 69	Theory and artifact	Prediction and design
Kaur, A., & Datta, A. (2015). A novel algorithm for fast and scalable subspace clustering of high-dimensional data. *Journal of Big Data*, *2*(1), 17	Artifact	Design
Khalilian, M., Mustapha, N., & Sulaiman, N. (2016). Data stream clustering by divide and conquer approach based on vector model. *Journal of Big Data*, *3*(1), 1	Artifact	Design
Nagwani, N. K. (2015). Summarizing large text collection using topic modeling and clustering based on MapReduce framework. *Journal of Big Data*, *2*(1), 6	Artifact	Design
O'Donovan, P., Leahy, K., Bruton, K., & O'Sullivan, D. T. J. (2015). An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *Journal of Big Data*, *2*(1), 25	Theory and artifact	Prediction and design
Pirouz, M., & Zhan, J. (2016). Optimized relativity search: Node reduction in personalized page rank estimation for large graphs. *Journal of Big Data*, *3*(1), 12	Artifact	Design
Prusa, J. D., & Khoshgoftaar, T. M. (2017). Improving deep neural network design with new text data representations. *Journal of Big Data*, *4*(1), 7	Artifact	Design
Sharma, S., & Toshniwal, D. (2017). Scalable two-phase co-occurring sensitive pattern hiding using MapReduce. *Journal of Big Data*, *4*(1), 4	Artifact	Design
Yang, Y., Zhang, K., Wang, J., & Nguyen, Q. V. (2015). Cabinet Tree: An orthogonal enclosure approach to visualizing and exploring big data. *Journal of Big Data*, *2*(1), 15	Artifact	Design
Young-Min, K. (2019). Feature visualization in comic artist classification using deep neural networks. *Journal of Big Data*, *6*(1), 56	Artifact	Design
Zhang, H., Raitoharju, J., Kiranyaz, S., & Gabbouj, M. (2016). Limited random walk algorithm for big graph data clustering. *Journal of Big Data*, *3*(1), 26	Artifact	Design
Brynjolfsson, E., Geva, T., & Reichman, S. (2016). Crowd-Squared: Amplifying the Predictive Power of Search Trend Data. *MIS Quarterly*, *40*(4), 941–962	Artifact	Design
Martens, D., Provost, F., Clark, J., & Junqué de Fortuny, E. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly, 40*(4), 869–888	Theory and artifact	Prediction and design

In this exercise, we differentiate between theory and artifact to highlight the nature of contribution, even though they are not mutually exclusive, and are often inseparable. It is evident from this sample that data science is frequently approached as a design study, rather than a research method or methodology. While the utility of data science contributions signifies the value of these studies to practice, there remains a realm of unseized opportunities for further theorization across different disciplines. Different factors may contribute to this status quo, including the shortage in guidelines for rigor in data science methodology, the low level of acceptance of such studies that do not "produce"

Table 8 Sample of studies using data science as a research method and object of study

Title of RMO-type studies	Contribution	Goal
Baechle, C., Agarwal, A., & Zhu, X. (2017). Big data driven co-occurring evidence discovery in chronic obstructive pulmonary disease patients. *Journal of Big Data*, *4*(1), 9	Theory and artifact	Explanation, prediction and design
Etani, N. (2015). Database application model and its service for drug discovery in Model-driven architecture. *Journal of Big Data*, *2*(1), 16	Theory and artifact	Prediction, design and action
Hayes, M. A., & Capretz, M. A. (2015). Contextual anomaly detection framework for big sensor data. *Journal of Big Data*, *2*(1), 2	Artifact	Design
Kumar, S., & Toshniwal, D. (2016). A novel framework to analyze road accident time series data. *Journal of Big Data*, *3*(1), 8	Theory and artifact	Prediction and design
Mavragani, A., & Tsagarakis, K. P. (2019). Predicting referendum results in the Big Data Era. *Journal of Big Data*, *6*(1), 3	Theory and artifact	Prediction and design
Subroto, A., & Apriyana, A. (2019). Cyber risk prediction through social media big data analytics and statistical machine learning. *Journal of Big Data*, *6*(1), 50	Theory and artifact	Explanation, prediction and design
Yang, J., & Yecies, B. (2016). Mining Chinese social media UGC: A big-data framework for analyzing Douban movie reviews. *Journal of Big Data*, *3*(1), 3	Artifact	Analysis and design
Liebman, E., Saar-Tsechansky, M., & Stone, P. (Forthcoming). The Right Music at the Right Time: Adaptive Personalized Playlists Based on Sequence Modeling. *MIS Quarterly, 43*(3), 765–786	Theory and artifact	Prediction, design and action
Son, J., Brennan, P.F., & Zhou, S. (Forthcoming). A Data Analytics Framework for Smart Asthma Management Based on Remote Health Information Systems with Bluetooth-Enabled Personal Inhalers. *MIS Quarterly, 44*	Artifact	Design and action
Mo, J., Sarkar, S., & Menon, S. (2018). Know When to Run: Recommendations in Crowdsourcing Contests. *MIS Quarterly*, *42*(3), 919–944	Theory and artifact	Prediction and design
Abbasi, A., Zhou, Y., Deng, S., & Zhang, P. (2018). Text Analytics to Support Sense-Making in Social Media: A Language-Action Perspective. *MIS Quarterly*, *42*(2), 427–464	Theory and artifact	Design
Lin, Y. K., Chen, H., Brown, R. A., Li, S. H., & Yang, H. J. (2017). Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach. *MIS Quarterly*, *41*(2), 473–496	Theory and artifact	Prediction and design
Zhang, K., Bhattacharyya, S., & Ram, S. (2016). Large-Scale Network Analysis for Online Social Brand Advertising. *MIS Quarterly*, *40*(4), 849–868	Artifact	Design

artifacts, the nature of knowledge and skills of researchers conducting data science studies, or the vague understanding of the nature of theoretical propositions that data science can enable producing. We argue for the latter and present some guidelines in the following section.

Discussion: data science as a research methodology

Here, we visit the data science process employed in research and practice, and recursively argue how theoretical development can take place along a systematic process of data analysis and knowledge extraction. The idea of traversing between data and theory is ancient, fundamental and not uncommon in other methodologies. Grounded

theory, for example, is data-driven and promotes for the inductive generation of theory from data [55]. Müller et al. [7] demonstrate parallels between data science studies and grounded theory in terms of understanding the data, the notion of being open to surprises in the data, and iteratively analyzing and searching for concepts. Similarly, inductive case study research—across different philosophical traditions—promotes for overlapping data collection, analysis and theorization [56–58].

In these guidelines, we focus on text analytics as a subdomain in data science. Text, and unstructured data in general, constitute the majority of the data we produce, both in practice and in research. Unstructured data (in part) sparked the rise of data science [29]. Key elements often discussed in text analytics are corpuses, documents, and words (or tokens). Three studies from three different disciplines, seeking three different theoretical claims from text analytics, are reviewed for illustrative purposes:

1. Study A: From information systems, Müller et al. [7] classify and predict helpfulness of online product reviews.
2. Study B: From service innovation and design, Antons and Breidbach [10] describe and analyze the extant literature in both areas, their overlaps, and delineate a research agenda for their joint future.
3. Study C: From economics, Bollen et al. [59] predict the stock market value based on public moods.

Research design

The importance of starting with a specific research question, problem, or opportunity in data-driven research has been highlighted previously in the literature [4, 60]. Like every research strategy, a research design is needed to ensure the planned research activities are sound, relevant and logical. A research design is simply a logical plan on how to answer the specific question and/or solve a problem. Accordingly, a good research design will help the researcher avoid problems such as collecting and analyzing data that do not address the initial question, or even answering the wrong questions. It is recommended that a DS research design specifies—at least—a research question, study goals, and unit of analysis.

The first step is to identify a (set of) research question(s) that the study aims to answer. Even though DS is a data-driven, theory-light, and primarily inductive approach, literature scanning and theoretical assessment are necessary to a) ensure the relevance of the research question, and b) justify the need for a DS methodology to address the particular question. Note that conceptual research questions and problems are different from their practical counterparts (see [47]) that may also present themselves in a data science study. While practical questions may be essential to ensure the relevance of the research, conceptual problems are important for theoretical contributions.

- A address the question of what makes an online review helpful.
- B address the question of what the existing body of knowledge on service innovation and service design entailed, and what may be possible future areas of integration and progress.

- C address the question of whether societies can experience mood states, as reflected in the public tweets, that would affect their collective decision making.

At the outset of the study, the research question is often generic. Yet it needs to be specific enough to allow for scoping and data collection. Study goals here refers to the theoretical goals that follow from the research question, which, according to Gregor [16], are description & analysis, explanation, prediction, and prescription (including design and action). Studies A and C both seek theoretical contributions that allow for prediction, while B seeks description and analysis of the status quo. Defining the unit of analysis also has implications on the following activities from sampling to preprocessing, analytics and interpretation. Especially in text analytics, it is beneficiary to clarify what unit of analysis does a "document" represent. For example, in our three exemplary studies, the document represents a product review, a research article, and a tweet, respectively. This ensures the following activities will enable the researcher to draw conclusions that relate to the research question.

Data collection

Data science is largely driven by the availability of data; that is datasets curated and opened up for public use. This availability expanded the possible research opportunities, but it is also associated with challenges if the data sources are not well justified provided the research problem—e.g. streetlight effect [4]. Collecting and curating large datasets may be consuming and requiring specific skills, but it may also be necessary for quality research. Being critical about the sources of data, its processes of production and intentions of publishing, is as important to ethical data studies as in other research strategies [11].

One strategy to address validity and reliability of a DS study is to perform triangulation; data triangulation is especially relevant in this step. Data triangulation is the use of multiple data sources that may represent different facets of the phenomenon [58]. In addition, thorough documentation of the data collection process that allows replicability is necessary to provide transparency [7]. This process includes tasks such as sampling and data reduction. The three studies conducted data collection to form their respective corpuses as follows:

- A downloaded a dataset of Amazon product reviews that is publicly available and curated earlier for research purposes. They applied sampling in two steps: (a) by product category to represent a specified unit of analysis, and (b) by excluding reviews with less than two helpfulness ratings in order to increase analytical reliability.
- B collected their dataset through publication database searches and aggregation. They documented their search steps thoroughly and followed established guidelines and benchmarks in their field.
- C collected two subsets of data that represent the study's different constructs: tweets to extract moods, and Dow Jones Industrial Average (DJIA) closing val-

ues for stock market indicators. They applied sampling on collected tweets using regular expressions to extract on tweets that have explicit emotional states.

Preprocessing

The specifics of preprocessing mainly depend on the chosen analytical technique and the overall research design. For instance, to cluster a corpus of papers, the researcher may need to represent the corpus in vector form, while to model topics this may be unnecessary. In text analytics, common preprocessing tasks include filtering out language- and context-specific stopwords, removing punctuation, reducing variability in words through stemming, lemmatization and/or case transformation [61]. The three cases follow this common preprocessing approach.

It is important to note that a preprocessing step in one study may be an analytics step in another, depending on the research design. To illustrate this, consider study C's research design, where sentiment analysis was applied as a preprocessing step to quantify the "public mood" construct, which is then used as an input variable to a neural network to build a predictive model. Sentiment analysis, in other studies, is used in the analytics step as the main analysis method [62]. Study C also conducted what is referred to as methodical triangulation; that is the use of multiple methods to measure a specific variable, which addresses the validity of such variable.

Analytics

This is the main step where insights and knowledge are extracted from data through the application of data mining, machine learning, and natural language processing (NLP) techniques, among others. In this section, we focus on text analytics and topic modeling. In addition to corpuses, documents and words, topic modeling brings another element into the picture: topics. Probabilistic topic modeling is an unsupervised technique that helps a researcher discover topics in a corpus of text based on the co-occurrence of terms in similar contexts. Topics denote an essential entity that binds documents with words, defined as "a distribution over a fixed vocabulary" [63, p. 78]. Essentially, the objective of topic modeling is to uncover topics expressed in a corpus in an automated approach.

Every technique is developed with a set of assumptions. Latent Dirichlet Allocation (LDA), an established topic modeling technique [64], has three key assumptions as follows:

(a) Topics exist independent of any data collection or generation,
(b) Documents and words are the only observed entities, and
(c) Topic structures, in terms of topics-document and document-word distributions, are hidden.

It is therefore important to reflect on similar assumptions in relation to the object of study, and the researcher's epistemological choices in order to avoid misfits and incoherence in resulting theoretical contribution. A researcher should motivate the choice of the specific technique in light of their study specifics and each technique's strengths and

weaknesses. For example, structural topic models (STM) may be best suited for political texts where metadata is of key importance [65], while understanding the change of topics over time may be best achieved through chronological topic modeling [66] or dynamic topic models [67]. Furthermore, each technique has a number of implementations (i.e. algorithms) available, which is important to communicate in the study. Assumptions of the technique as well as its implementation extend to the assumptions behind and boundaries of the proposed theoretical claim.

The three studies develop their models as follows:

- A employs topic modeling for feature (variable) selection; meaning that they provide each review with a probability (weight) for each of the top identified topics. Along with other variables from the literature, they use random forests to classify and predict the helpfulness of the review.
- B use topic modeling to extract, label and classify different topics in research texts. In order to go from state-of-the-art towards future trajectories, they computed a linear trend of topics over time, as well as network structure between topics.
- C developed the hypothesis that including the public mood measure into existing stock market prediction models will enhance its accuracy. The new model was developed through a self-organizing fuzzy neural network.

In all three studies, multiple DS methods are used to answer their research question. This is another way of using methodical triangulation to enhance the research design. Documenting all the parameters that led to such results is crucial for the transparency and replication of the study. Three levels of evaluation of the results are recommended: a) evaluation metrics that follow from the chosen technique; e.g. predictive error, or cluster coherence, b) comparative performance of the resulting model with accepted benchmarks or baseline models, and c) utilizing human coders (e.g. expert panel) to provide input on the generated insight.

Interpretation and theorizing

This step is concerned with converting the extracted knowledge to a theoretical contribution adding to the body of knowledge. A common challenge in interpretation of DS studies is the comprehensibility of the results [4, 7]. This challenge is amplified with specific techniques that are regarded as black boxes or have a high predictive and low explanatory power. Thus, the precision paradox emerges here, and the study goals need to be aligned. In case the study goal is primarily predictive, further analysis may be required to enrich the interpretation of the model to be able to generate theoretical propositions. On the other hand, some techniques are set out to open up the so-called black boxes. For instance, a key computational problem with topic modeling is identifying the hidden structures that likely generated the observed corpus. The scholarly challenge here is interpreting this generative process (represented by probability distributions of topics) and providing it with context that allows for theoretical contribution.

What follows this localized interpretation is examining the results against the existing body of knowledge to assess its relevance. Theoretical triangulation is recommended in this step; that is interpreting the results in light of existing theory or benchmarking

against existing artifacts [7]. Human intervention, particularly other researchers, would enhance the reliability and robustness of the results.

Since this paper is primarily concerned with mid-range theories, patterns in data alone do not suffice for theoretical contribution. It is rather expected to specify propositions that provide context in the subject matter and highlight key concepts addressed in the study [18, 27]. Matching emerging insights and relations to theoretical concepts and propositions, respectively, allows lifting the results into the theoretical domain. It is often through iterations between the different study phases—especially analytics and interpretation—that theoretical saturation is achieved, and sound theoretical propositions are formed.

What constitutes a contribution could be discussed lengthily, but for the purpose of this paper we share the views of Hassan and Lowry [18, p. 11]:

> *"It requires an understanding of the theory's originality, how it modifies the rules of discourse in the field, what hidden assumptions underlie the theory, what new concepts are being introduced and how they impact the discourse, what laws will be affected or constructed as a result, and the range and scope the theory is expected to cover."*

To summarize, we revisit the three exemplary studies with regards to their interpretation and theoretical contributions:

- A provided two types of relational propositions: the key constructs that are predictive of a review helpfulness, and the direction of correlation (positive or negative) between the independent and dependent variables. The latter proposition, deemed necessary for explanatory interpretation, was developed through a second iteration analytics and interpretation.
- B also provided two propositions: definitional propositions represented by the extracted topics, and relational propositions represented by the identified topic network including topic nodes and edges. Theorization was also extended to identify mechanisms for future research that would utilize the domains' trajectories.
- C developed the proposition in the form of a testable hypothesis including the constructs matching the independent variables identified through semantic analysis.

This discussed methodology and constituent guidelines can be summarized in Fig. 1.

Example: exploring data-driven innovation through text analytics

In order to further illustrate using data science for theoretical development, we adopt it as a methodology to understand digital innovation.

Research design

This study started out with the exploratory question of how data-driven innovations emerge in open ecosystems such as smart cities. Innovations emerging in open ecosystems are particularly interesting due to the distribution of actors and control, which affects where and how key innovation activities along the innovation process take place. The study was motivated by two problems. First, there is the conceptual problem of the

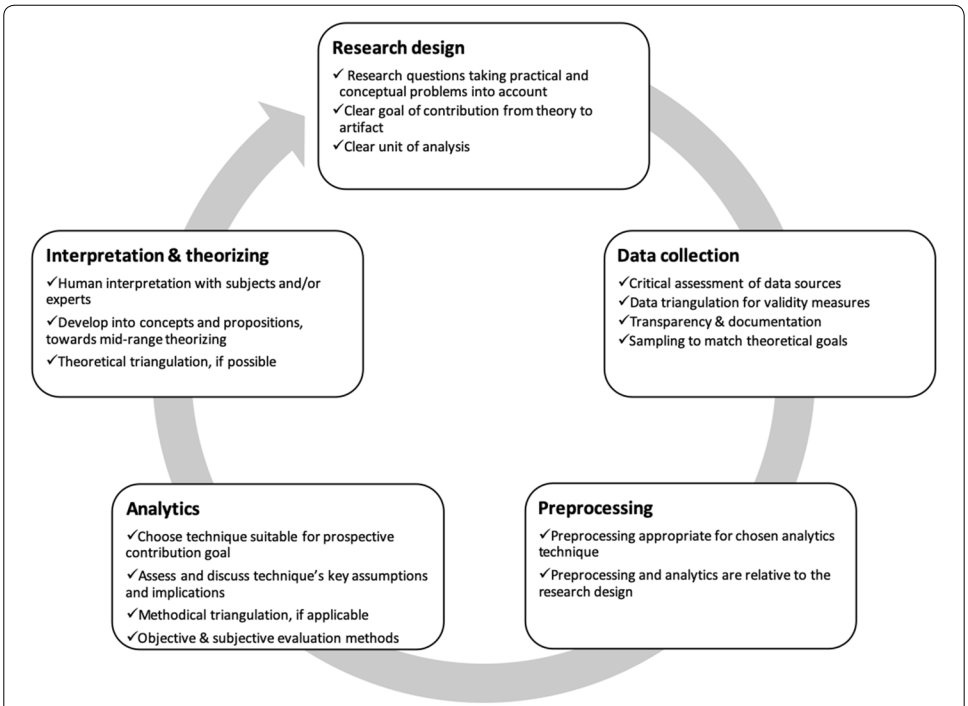

Fig. 1 Data science as a methodology for theoretical contribution

lack of understanding of data-driven innovation as a process. Second, the amount of funding that goes into innovation programmes—e.g. European Commission is spending around EUR 80 billion over 7 years on innovations from idea to market [68]—is significant with few innovations actually commercializing. Accordingly, the goal of the study was to make a relational proposition of the process type [69] with a key focus on data-driven activities and events. Thus, the main unit of analysis are these activities and events. While the problem can be addressed qualitatively to inductively uncover such process, the amount of documentation that was deemed relevant was challenging to navigate. Hence, a text analytics approach was required.

The research setting includes studying four different data-driven innovations that have taken shape within the context of smart cities. This study is part of a research and innovation project funded by the European Commission—OrganiCity—from 2015 till 2018. The scope of the project was experimentation with IoT, urban data and co-creation tools in three smart cities: London, Santander and Aarhus. The project operated by providing the experimentation facility and offering funding to teams that are interested in innovating with data, IoT, and other digital tools to design and deliver their own data-driven solutions. A total of 37 teams have been funded by the project over 2 open calls and a period of execution of 6 months each. Each team was funded with up to EUR 60,000 per call.

Sampling and data collection

The cases were selected based on theoretical sampling to achieve a diversity of how data and analytics are used in relation to innovation outcomes and activities [70]. In addition, the cases were selected operating in the same city (London), with similar team

Table 9 Colour-in city data sources

Case	Data sources	Documentation
Colour-in city	2 interviews; 3 team members 3 reports 6 blog entries 12 design tools Documentation of 3 visual artifacts	111 min; 20 pages 46 pages 39 pages 62 pages 9 pages

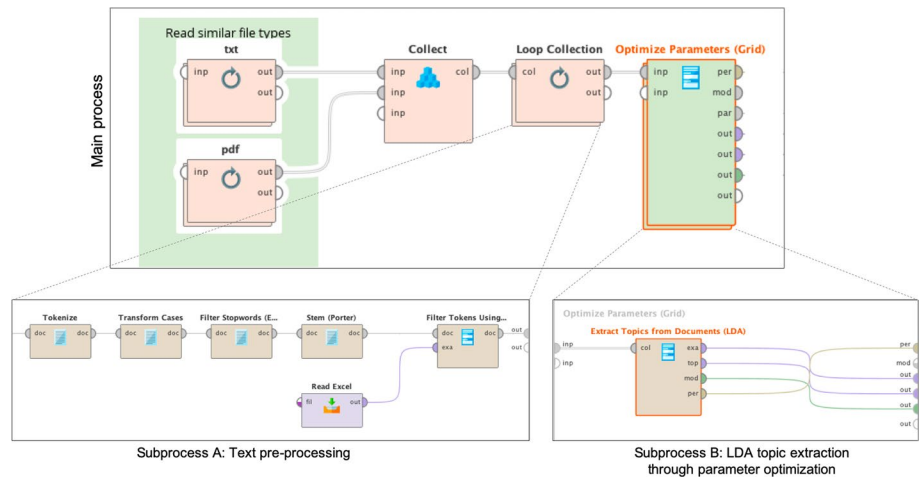

Fig. 2 RapidMiner process and subprocesses

sizes, technological maturity, funding amounts, and operating conditions. For each case, both primary and secondary research data were collected. Primary data constituted in-depth interviews (both individual and group) with the respective teams. Secondary data included progress reports, funding contracts and planning documents, public blogs and other communications within the project consortium. In total, the corpus contained 160 documents that were primarily produced by the team to report on their journey.

It is important to note that all text was produced within the funding mechanism, meaning that some activities and events may have been highlighted more than others in public communications. Accordingly, we address this using data triangulation where internal documentation and interviews were used to capture challenges and undesirable events. All data was converted to textual format and combined in a single repository, either in.txt or.pdf format.

The next three phases of research were conducted five times: one on single case's set of documents to draw case-specific conclusions, as well as on the whole corpus to draw cross-case conclusions. In what follows we describe the specifics of just one case for illustration: the case of developing a chatbot for understanding over crowdedness through subjective and objective wellbeing data [71]. Table 9 presents an overview of the case data sources.

Text preprocessing

In this study, all preprocessing and analytics is conducted using RapidMiner version 9.3. RapidMiner is a platform written in Java that provides an integrated development environment for different data science activities. The overall process diagram is shown in Fig. 2. Each box represents an operator performing a specific task or subprocess. In the top part of the figure lies the main process, in which the respective.txt and.pdf files are read as input.

In order to extract topics using LDA, the corpus needed to be represented as a collection of documents (in RapidMiner's terms, that is an Object Collection). Thus, the repository was scanned through a Loop Files operator, in which each file was read as a RapidMiner Document. Then a Collect operator was applied to create an Object Collection before this collection was looped to perform the following tasks (Subprocess A):

- Tokenize: Transforms the text of a document into a sequence of 1−n tokens.
- Transform cases: Transforms all tokens to lower case.
- Filter stopwords: Removes tokens that match a built-in English stopwords (e.g. and, an).
- Stem: Transforms all words to their stem/origin (i.e. define, defined, defining—all become *defin*).
- Filter using a wordlist: Removes tokens that match a provided list (represented by the "Read Excel" operator). This step was introduced after certain tokens were found to skew the resulting topics. Accordingly, a list of those tokens was compiled and used. The list is case-specific and contained tokens not captured by the built-in list of stopwords (e.g. yeah, etc., http, www, com), as well as entity names that were skewing the word distribution (e.g. OrganiCity, team names).

Text analytics

This step is represented in subprocess B as shown in Fig. 2. In this step, RapidMiner's native LDA operator is applied, which is based on Newman et al.'s [72] LDA implementation of parallel topic modeling. The LDA operator was applied iteratively within an optimization operator, using the number of topics and corresponding performance measures to select a model. Results have shown the best cut-off to be 10 topics, balancing between the performance metrics (i.e. Perplexity of 446.4 and log likelihood score of −95,046.4) and number of topics that can be reasonably interpreted. When proposing a new algorithm, these metrics are compared against published thresholds that other algorithms achieve against the same corpus. However, since we are using an established algorithm in this example, we rather rely on the teams' and experts' interpretation to evaluate the coherence of the topics.

After initial screening, one topic was excluded due the high weights of reporting terminology used within the consortium. After examining the remaining 9 topics, they were visualized using word/tag clouds. Figure 3 below shows the word clouds used with the included terms and their represented weights. These visualizations

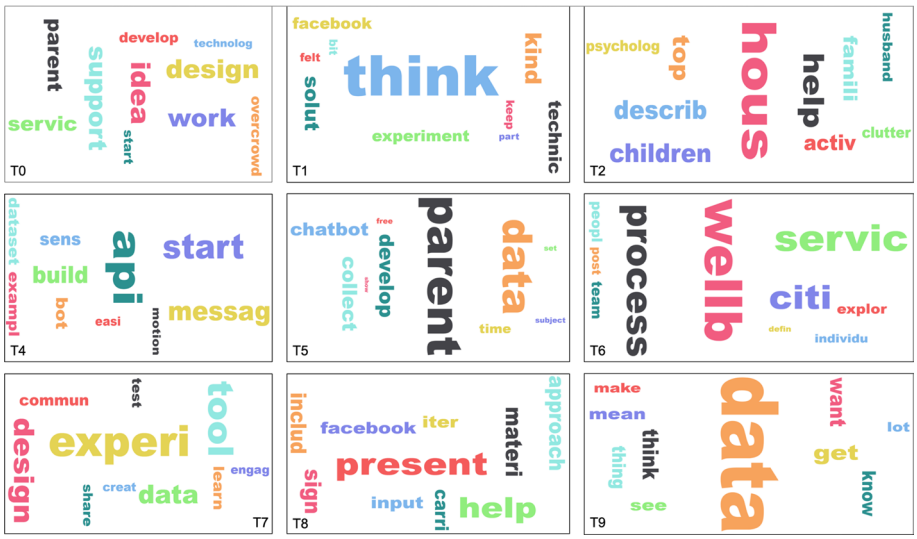

Fig. 3 Word clouds for Colour-in City's LDA topics used for evaluation and interpretation

Table 10 Topics' different codes based on multiple rating

T_ID	Expert panel	Innovation team	Researchers' interpretation
T0	Problem formulation	Design ingredients (process)	Design process
T1	Front end of innovation	Defining solution in uncertainty	Fuzzy front end (FFE)
T2	Need finding	Describing the user/audience	User understanding
T3	Questionnaire	Reporting (interim survey)	Reporting jargon
T4	Technical implementation	Data infrastructure	Data infrastructure
T5	Service end-user interaction	Chatbot function	User interaction
T6	Stakeholder analysis	Impact on each stakeholder	Stakeholders
T7	Experimenting	Approaches to innovation	Methods/approaches
T8	Communication of innovation	Usefulness of chatbot	Presenting innovation
T9	Data analysis	Insights using analytics	Data analytics

were then used for interpretation by two different groups: a panel of three expert researchers with focus on digital innovation, and a workshop with two of the case's team members. Each session lasted between 30 and 60 min. Table 10 shows each of the extracted topics and each group's interpretation. The discrepancy shown in T0 is mainly caused by the lack of context knowledge on the panel's part, which still leads to 90% inter-rater agreement, which is above the threshold (70%) established in other methodologies [73].

Theoretical interpretation

The identified topics denote salient concepts that the innovation teams have communicated through text. However, these concepts relate to different phases of the digital innovation process, which was not evident from topic modeling alone. Accordingly, the existing body of knowledge on the digital innovation process was used as a theoretical lens to interpret those concepts in relation to the four digital innovation phases:

ideation, development, diffusion, and impact. This process was challenged in two ways through this analysis.

On one hand, specific topics highlighted activities that require a new distinct phase pertaining to data-driven innovations. After examining the FFE concept (T1), it was clear that the included events and activities neither exclusively belong to the ideation, nor to the development of digital innovations. Rather, it is a distinct phase in between where ideas undergo critical examination through matching them with potential resources—from financial resources (e.g. funding and investment) to available datasets that would drive the innovation. This phase of critical examination is not new, but was considered part of the ideation phase [74]. However, the process that such data-driven innovation went through included extensive scanning and assessment efforts with elements from ideation as well as experimentation that indicated a distinct innovation phase.

On the other hand, the stage gate nature of the process was challenged when the topics have shown activities that are difficult to place in a linear order. The teams affirmed that a few of those activities (e.g. design ingredients and user interactions) have been iteratively conducted more than once while iterating between the innovation phases. These challenges could be attributed to the nature of digital innovation where data is the key resource, or the openness of the ecosystem leading to more opportunities and challenges than within a single firm. In addition, the extracted topics shed light on specific concepts that are not typically addressed in digital innovation: FFE is rather investigated in product and process innovations.

Matching the emerging concepts with this process lifts them to the theoretical domain, while staying grounded with innovation events and activities that are directly observable. In this study, a middle-range process-type proposition is sought, one that is bounded within digital innovation. The study contributes to the existing body of knowledge in the following ways [19]:

- The *what*: through acknowledging a distinct phase of critical examination of the innovation that has been previously overlooked or downplayed.
- The *how*: through stressing on the iterative rather than linear nature of the process.
- The *why*: through calling for new logic underlying the process that accommodates for (a) the nature of data-driven innovations, and (b) the blurring boundaries between innovation teams and their users.

The study presented in this example has also various practical implications. It views data-driven innovation as a process journey that requires navigation rather than control [75], providing innovators with insights on phases that could be more challenging than others. Iteration will also encourage innovators to work with faster and cheaper iterations and hands-on data experimentation instead of investing in data infrastructures that may be irrelevant in later iterations.

Conclusion

The organizational science has been criticized for the lack of innovative research due to the dominance of gap-spotting type of research [14]. Similarly, concerns voicing the need to develop forward thinking and innovative research has drawn the attention of

the IS Scholars [76, 77]. In order for us to be able to overcome the problem, in organizational sciences, Alvesson and Sandberg [14] suggested moving away from the boxed-in towards box-breaking research. Bringing theoretical contributions to the IS field via data science is, we believe, a type of box-breaking research.

Data science has been considered as a wave that brings a plethora of opportunities to scientific research [3, 36], and IS discipline is no exception. This wave coincides with recurring concerns over theorizing practices in IS, and how the excessive borrowing of theories and philosophies from other domains results in an incoherent body of knowledge. Recommendations to overcome this problem include following more inductive data-driven approaches with less weight given to pre-existing theory, seeking middle-range theories, and focusing on concepts [15, 18].

In this paper, we focused on these recommendations and presented how data science can be used as a methodology to build theoretical contributions. We specifically focused on extracting knowledge from textual data, from setting a research design onto theorization with particular attention to concepts and propositions. While the key issue addressed in this paper is one of foundational theorizing practices in information systems e.g. see [27] in light of leveraging data science as a research methodology, it also highlights how this approach enables novel and varied practices of theorizing.

However, this was not without its challenges, such as separating data science as a method from the object of the study (e.g. studies developing data analytics frameworks). This study is also limited by the scope of the two journals selected to represent IS research on analytics and data science. It is also focused on text analytics. Further research is needed to examine other outlets, analysis methods and theory types, such as process type theories. The illustrative example needs to be presented with evidence from the four cases to draw more accurate conclusions on data-driven innovation. It is also beneficial to compare DS and other methodologies in answering similar research questions and compare the findings.

Abbreviations

DS: data science; IS: information systems; LDA: Latent Dirichlet Allocation; JBD: Journal of Big Data; MISQ: Management Information Systems Quarterly; IoT: Internet of Things; IT: information technology; FFE: fuzzy front end (of innovation); DJIA: Dow Jones Industrial Average.

Acknowledgements

Not applicable.

Authors' contributions

Both authors worked on the conception and design of the paper. The first author conducted the review and data collection for the provided example. Both authors agreed on the method and collaboratively interpreted the results. Both authors read and approved the final manuscript.

Funding

Open access funding provided by Lulea University of Technology. The provided example is an outcome from a study partially funded by the OrganiCity project, funded by European Union's Horizon 2020 research and innovation program under the Grant Agreement No. 645198

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 October 2019 Accepted: 23 December 2019
Published online: 09 January 2020

References

1. R. Agarwal and V. Dhar, Editorial—Big data, data science, and analytics: the opportunity and challenge for IS research. INFORMS. 2014.
2. Cukier K. Special report: data, data everywhere. The Economist. 2010.
3. Hey AJ, Tansley S, Tolle KM. The fourth paradigm: data-intensive scientific discovery, vol. 1. Redmond: Microsoft Research; 2009.
4. Elragal A, Klischewski R. Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. J Big Data. 2017;4(1):19.
5. Frické M. Big data and its epistemology. J Assoc Inf Sci Technol. 2015;66(4):651–61.
6. Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data Soc. 2014;1(1):1–12.
7. Müller O, Junglas I, vom Brocke J, Debortoli S. Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur J Inf Syst. 2016;25(4):289–302.
8. Kelling S, et al. Data-intensive science: a new paradigm for biodiversity studies. Bioscience. 2009;59(7):613–20.
9. Kitchin R. The data revolution: big data, open data, data infrastructures and their consequences. Thousand Oaks: Sage; 2014.
10. Antons D, Breidbach CF. Big data, big insights? Advancing service innovation and design with machine learning. J Serv Res. 2018;21(1):17–39.
11. Neff G, Tanweer A, Fiore-Gartland B, Osburn L. Critique and contribute: a practice-based framework for improving critical data studies and data science. Big Data. 2017;5(2):85–97.
12. Latour B. Tarde's idea of quantification. na, 2010.
13. Berente N, Seidel S, Safadi H. Research commentary—data-driven computationally intensive theory development. Inf Syst Res. 2018;30(1):50–64.
14. Alvesson M, Sandberg J. Has management studies lost its way? Ideas for more imaginative and innovative research. J Manag Stud. 2013;50(1):128–52.
15. Grover V, Lyytinen K. New state of play in information systems research: the push to the edges. MIS Q. 2015;39(2):271–96.
16. Gregor S. The Nature of Theory in Information Systems. MIS Q. 2006;30(3):611–42.
17. Lee AS. Editor's comments: research in information systems: what we haven't learned. MIS Q. 2001;25(1):v.
18. Hassan NR, Lowry PB. Seeking middle-range theories in information systems research. In: International conference on information systems (ICIS 2015), Fort Worth, TX, December, 2015. p. 13–8.
19. Whetten DA. What constitutes a theoretical contribution? Acad Manag Rev. 1989;14(4):490–5.
20. Merriam-Webster. Definition of Concept. 2019. https://www.merriam-webster.com/dictionary/concept. Accessed 04 July 2019.
21. Dubin R. Theory building. Mumbai: Free Press; 1969.
22. Bacharach SB. Organizational theories: some criteria for evaluation. Acad Manag Rev. 1989;14(4):496–515.
23. Gibbs JP. Sociological theory construction. Hinsdale: Dryden Press; 1972.
24. Hevner A, March ST, Park J, Ram S. Design science research in information systems. MIS Q. 2004;28(1):75–105.
25. Sein M, Henfridsson O, Purao S, Rossi M, Lindgren R. Action design research. MIS Q. 2011;35(1):37–56.
26. Lee AS, Baskerville RL. Generalizing generalizability in information systems research. Inf Syst Res. 2003;14(3):221–43.
27. Hassan NR, Mathiassen L, Lowry PB. The process of IS theorizing as a discursive practice. J Inf Technol. Forthcoming. 2019.
28. Orlikowski WJ, Iacono CS. Research commentary: desperately seeking the 'IT' in IT research—a call to theorizing the IT artifact. Inf Syst Res. 2001;12(2):121–34.
29. Dhar V. Data science and prediction. Commun ACM. 2013;56(12):64–73.
30. Kelleher JD, Tierney B. What is data science?. In: Data Science, MIT Press; 2018. p. 1–38.
31. Ahalt S. Why Data Science?. In: Presented at the National Consortium for Data Science. Chapel Hill; 2013.
32. Merriam-Webster. Definition of Scientific method. 2019. https://www.merriam-webster.com/dictionary/scientific+method. Accessed 08 July 2019.
33. Orlikowski WJ, Baroudi JJ. Studying information technology in organizations: research approaches and assumptions. Inf Syst Res. 1991;2(1):1–28.
34. Hassan NR, Mingers J, Stahl B. Philosophy and information systems: where are we and where should we go? Eur J Inf Syst. 2018;27(3):263–77.
35. Myers MD. Qualitative research in information systems. Manag Inf Syst Q. 1997;21(2):241–2.
36. Bell G, Hey T, Szalay A. Beyond the data deluge. Science. 2009;323(5919):1297–8.
37. Kuhn TS. The structure of scientific revolutions. Chic Lond. 1962.
38. Lenca P, Petit J-M. Guest editor's introduction: special issue on discovery science 2012. J Intell Inf Syst. 2015;44(2):191–2.
39. Andrejevic M. Infoglut: How too much information is changing the way we think and know. Abingdon: Routledge; 2013.
40. Goes PB. Editor's comments: big data and IS research. MIS Q. 2014;38(3):iii–viii.
41. Rai A. Editor's comments: synergies between big data and theory. MIS Q. 2016;40(2):iii–ix.
42. Saar-Tsechansky M. The business of business data science in IS journals. MIS Q. 2015;39(4):iii–vi.
43. Debortoli S, Müller O, Junglas I, vom Brocke J. Text mining for information systems researchers: an annotated topic modeling tutorial. Commun Assoc Inf Syst. 2016;39:110–35.

44. Trifunovic N, Milutinovic V, Salom J, Kos A. Paradigm shift in big data supercomputing: dataflow vs controlflow. J Big Data. 2015;2(1):4.
45. Bughin J. Big data, big bang? J Big Data. 2016;3(1):2.
46. Gregor S, Hevner AR. Positioning and presenting design science research for maximum impact. MIS Q. 2013;1:337–55.
47. Geva T, Oestreicher-Singer G, Efron N, Shimshoni Y. Using forum and search data for sales prediction of high-involvement projects. MIS Q. 2017;41(1):65–82.
48. Goswami K, Park Y, Song C. Impact of reviewer social interaction on online consumer review fraud detection. J Big Data. 2017;4(1):15.
49. Mavragani A, Ochoa G. Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. J Big Data. 2018;5(1):30.
50. Sohangir S, Wang D, Pomeranets A, Khoshgoftaar TM. Big data: deep learning for financial sentiment analysis. J. Big Data. 2018;5(1):3.
51. Agarwal A, Baechle C, Behara RS, Rao V. Multi-method approach to wellness predictive modeling. J Big Data. 2016;3(1):15.
52. Asri H, Mousannif H, Al Moatassime H. Reality mining and predictive analytics for building smart applications. J Big Data. 2019;6(1):66.
53. Mo J, Sarkar S, Menon S. Know when to run: recommendations in crowdsourcing contests. 2018.
54. Abbas A, Zhou Y, Deng S, Zhang P. Text analytics to support sense-making in social media: a language-action perspective. MIS Q. 2018;42(2):427–64.
55. Glaser BG, Strauss AL. The discovery of grounded theory: strategies for qualitative research. Piscataway: Transaction Publishers; 2009.
56. Eisenhardt KM. Building theories from case study research. Acad Manag Rev. 1989;14(4):532–50.
57. Walsham G. Interpretive case studies in IS research: nature and method. Eur J Inf Syst. 1995;4(2):74.
58. Yin RK. Case study research: design and methods. Thousand Oaks: Sage publications; 2013.
59. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci. 2011;2(1):1–8.
60. Booth WC, Colomb GG, Williams JM. The craft of research. Chicago: University of Chicago Press; 2003.
61. Blei DM, Lafferty JD. A correlated topic model of science. Ann Appl Stat. 2007;1(1):17–35.
62. Dong R, O'Mahony MP, Schaal M, McCarthy K, Smyth B. Combining similarity and sentiment in opinion mining for product recommendation. J Intell Inf Syst. 2016;46(2):285–312.
63. Blei D. Probabilistic topic models. In: Proceedings of the 17th ACM SIGKDD international conference tutorials. 2011. p. 5.
64. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res. 2003;3:993–1022.
65. Roberts ME, et al. Structural topic models for open-ended survey responses. Am J Polit Sci. 2014;58(4):1064–82.
66. Masada T, Takasu A. ChronoSAGE: diversifying topic modeling chronologically. In: International conference on web-age information management. 2014. p. 476–9.
67. Blei DM, Lafferty JD. Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning. 2006; p. 113–20.
68. European Commission. Funding for innovation. Internal market, industry, entrepreneurship and SMEs. July 05 2016. https://ec.europa.eu/growth/industry/innovation/funding_en. Accessed: 15 July 2019.
69. Van de Ven AH, Huber GP. Longitudinal field research methods for studying processes of organizational change. Organ Sci. 1990;1(3):213–9.
70. George G, Lin Y. Analytics, innovation, and organizational adaptation. Innovation. 2017;19(1):16–22.
71. Colour-in City, http://colourincity.com/. 2017. http://colourincity.com/. Accessed: 20 Mar 2017.
72. Newman D, Asuncion A, Smyth P, Welling M. Distributed algorithms for topic models. J Mach Learn Res. 2009;10:1801–28.
73. Miles MB, Huberman AM. Qualitative data analysis: an expanded sourcebook. Thousand Oaks: Sage; 1994.
74. Garud R, Gehman J, Kumaraswamy A, Tuertscher P. From the process of innovation to innovation as process. In: The SAGE Handbook of Process Organization Studies. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd; 2016. p. 451–65.
75. Van de Ven AH. The innovation journey: you can't control it, but you can learn to maneuver it. Innovation. 2017;19(1):39–42.
76. Agarwal R, Lucas HC Jr. The information systems identity crisis: focusing on high-visibility and high-impact research. MIS Q. 2005;29(3):381–98.
77. Klein HK, Hirschheim R. The structure of the IS discipline reconsidered: implications and reflections from a community of practice perspective. Inf Organ. 2008;18(4):280–302.

Publisher's Note