BIG DATA-DRIVEN THEORY BUILDING: PHILOSOPHIES, GUIDING PRINCIPLES, AND COMMON TRAPS

Arpan Kumar Kar

Department of Management Studies, Indian Institute of Technology Delhi, India

Spyros Angelopoulos

Durham University Business School, Durham University, United Kingdom

H. Raghav Rao

College of Business, University of Texas at San Antonio, USA

Abstract: While data availability and access used to be a major challenge for information systems research, the growth and ease of access to large datasets and data analysis tools has increased interest to use such resources for publishing. Such publications, however, seem to offer weak theoretical contributions. While big data-driven studies increasingly gain popularity, they rarely introspect why a phenomenon is better explained by a theory and limit the analysis to data descriptive by mining and visualizing large volumes of big data. We address this pressing need and provide directions to move towards theory building with Big Data. We differentiate based on inductive and deductive approaches and provide guidelines how may undertake steps for theory building. In doing so, we further provide directions surrounding common pitfalls that should be avoided in this journey of Big-Data driven theory building.

Keywords: Big data; Information systems; Artificial intelligence; Machine learning; Theory building; Computational social science.

1. INTRODUCTION

The availability of Big Data for scientific research has increased in recent times, as digital transformation (DT) initiatives are maturing globally, assisted by the growth of computational capabilities (Angelopoulos et al., 2023; Grover et al., 2020; Struijk et al., 2023). While data availability and access used to be a major challenge for information systems (IS) scholars, the increasing availability in volume, velocity, and variety of data has resolved this considerably. Due to such growth and ease of access to large datasets, there has been a rush to use them for research (Kar & Dwivedi, 2020). Further access to data from social media platforms has increased such propensity to undertake Big Data driven research. However, more often than not, such studies seem to offer weak theoretical contributions to the IS field (Struijk et al., 2022), while the need for practical contributions is becoming increasingly apparent (Davison, 2022). Big Data driven studies are increasingly gaining popularity within the broader IS field, they rarely, however, introspect why a phenomenon could be better explained by a theory and limit their analysis to what is happening by merely mining and crunching large volumes of data. Such studies tend to collect data and showcase applications of advanced algorithmic solutions for the visualization of large volumes of data by demonstrating

computational techniques. Oftentimes, however, such studies fail to make any contribution to the theoretical context within which the problem is situated and, thus, would lack generalizability as well as causal inferences, and may end up not holding the test of time. Such studies do not attempt to explain why a particular phenomenon is witnessed and the data descriptions rarely contribute towards theory building. Such studies, therefore, tend to have a very lose connection with the relevant theories and information technology (IT) artefacts and do not attempt to elucidate causality, paying unnecessary attention to the data collection and analysis, while since the data is often dated, such studies lose timeliness (Grover et al., 2020).

Our special issue aims to guide IS scholars to develop impactful research based on Big Data, while making strong theoretical contributions. Research directions provided in Kar and Dwivedi (2020) are used to guide research towards contributing in IS theory. In this editorial, we attempt to address the following guiding questions and provide directions for future research:

1. What is Big Data driven research?
2. How should researchers plan to build theory in Big Data driven studies?
3. How should researchers avoid common traps in Big Data driven research?

In doing so, we extend the discussions in Kar and Dwivedi (2020) and Miranda et al. (2022) to provide directions for Big Data driven research. In the following sections, we discuss how Big Data driven research has evolved, and provide guidelines for inductive as well as deductive Big Data driven research. We also provide examples of multiple studies that have been published in established journals of the field, which have followed similar guiding principles. This is followed by a discussion on traps to avoid in Big Data driven research. In the penultimate section, we discuss the articles of the special issue, and we conclude by delineating an agenda for future research on the topic.

2. HOW BIG IS BIG DATA?

Whilst the contemporary IS literature tends to emphasize the challenges associated with the growing complexities and volume of Big Data, it is important to remember that such issues have long been a topic of discussion in the Computer Science community for over half a century. In fact, one of the most important venues in the field, the VLDB (Very Large DataBases) conference, was established back in 1975 (Kerr, 1975), and the first proceedings from this conference highlight the focus on similar concerns that continue to be relevant in current debates surrounding Big Data, both within academia as well as the industry at large.

The concept of 'Big Data' typically refers to large volumes of data that require expensive IT infrastructure. Some, however, argue that 'Big' is a relative term and what was once only possible with mainframe computers can now be accomplished using standard software on a desktop computer (Manovich, 2011). Others point out that if 'Big Data' simply meant a large quantity of data, we would call it "Lots of Data" (Williams, 2012). Some definitions of the concept incorporate the required capacity to process the data, such as McKinsey Global Institute's definition of "*datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze*" (Manyika et al., 2011). In their definition, Gartner introduced the end-user perspective and the time dimension, defining Big Data as data that "*exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population*" (Merv, 2011). While Gartner is known for coming up with the dimensions of *volume*, *velocity*, and *variety* as defining for Big Data, these factors alone do not capture the distinct challenges they pose.

Concurrently, 'Big Science' is a well-known source of Big Data, which is characterized by the production of large volumes of data that require significant storage as well as computational resources and represents unique opportunities to investigate timely and topical IS elements. For instance, Venters et al. (2014) conducted research on digital coordination using data from an experiment at CERN that generated "an equivalent to 15 million gigabytes of data per year or a DVD every 5 seconds" (Venters et al., 2014, p. 933). To analyse such data, scientists at CERN employed 150,000 computers distributed across 600 sites in 62 countries. In another notable example, Chen et al. (2012) refer to the Sloan Digital Sky Survey, which, during its ten years of operation, collected data at the rate of 200 gigabytes per night and created 3D maps containing over 930,000 galaxies and 120,000 quasars, covering more than a quarter of the night sky. Compared to 'Big Science', the demands and complexities of organizational datasets may seem modest, but there are still examples, such as organizations providing popular social networking services, that pose significant challenges (Angelopoulos et al., 2023; Georgiadou et al., 2020). While size is often considered the primary concern, the challenges associated with Big Data in cyber-social systems can be far more complex. Such challenges include the loss of original contextual information when combining datasets from various sources (Angelopoulos et al., 2023), the intricate nature of socially-situated systems-in-use (Struijk et al., 2020; 2023), which includes context-dependent relevance and immediacy of information (Price et al., 2015), as well as the behavioural inferences drawn from the use of digital technologies and the ethical as well as legal dilemmas that surround the collection and use of sensitive data for purposes that may not have been envisioned (Angelopoulos et al., 2021).

While the concept of Big Data encapsulates a multidimensional and multiscale nature (Lazer et al., 2009), it remains a *sociocultural* phenomenon encompassing large datasets and the associated procedures for manipulating and analysing them, representing a paradigm shift in research, and thinking (boyd & Crawford, 2012). Such datasets can generate higher levels of intelligence and knowledge, leading to novel insights that were previously impossible, and that are perceived as objective and accurate. There is a need, thus, for accuracy of algorithmic solutions and computational resources for optimizing the collection, and analysis of large data sets, with the goal of identifying patterns that lead to economic, social, technical, and legal claims (Angelopoulos et al., 2023; 2021). Big Data, thus, is not solely about the mere size of the data, but rather, its capacity to search, aggregate, and cross-reference such data to generate fresh insights, ultimately for contributing to the generation and advancement of theory.

It is important to note that this special issue focuses solely on the opportunities and challenges of theory building with Big Data. However, it is worth mentioning that 'Big Compute' is also a related concept to our discussion but falls outside our scope. Big Compute deals with the challenges that arise from using computationally intensive methods in fields such as genomics (e.g., Marx, 2013), weather modelling (e.g., Malakar et al., 2013), and healthcare (Heising & Angelopoulos, 2021, 2022). While there may be some overlap in how Big Data and Big Compute address certain challenges, Big Data is primarily concerned with the exploration of large, complex datasets across multiple levels and dimensions, whereas Big Compute is primarily focused on optimizing hardware as well as software performance.

3. THEORY BUILDING IN INFORMATION SYSTEMS RESEARCH

Broadly speaking, the theoretical underpinnings of IS research tend to come mainly from behavioural theory, management theory, organization theory, computer science theories, and systems theory (Barki et al., 1993). Apart from the core computer science theories, the other related theories enable IS scholars to explain how users interact with technology artifacts

within individual, organizational, social, and political contexts, and the impact of such interaction. Theory building, however, seems to have been disrupted by the current trends in big data-driven research, whereby the essence of contributing to theory is increasingly seen to be lacking at all levels of analysis. Concurrently, Big Data driven research may inspire contributions towards design science and action research, whereby innovative solutions may also be created which help to define ideas, capabilities, practices, and innovative products or services through big data analysis (Angelopoulos et al., 2021; Hevner et al., 2004). The core of IS research is conceptualized through topics essential to the broader discipline, focusing on the sociotechnical capabilities, practices, and behaviours involved in planning, designing, constructing, implementing, and using Information Technology (IT) artifacts by different stakeholders (Benbasat & Zmud, 2003). The focus may be towards thematic areas surrounding IT and organizations, IS development, IT, and individuals, IT and markets, IT and groups (Sidorova et al., 2008). Recent reviews in analysing the nomological network of IS literature indicate there are well defined thematic areas like e-commerce and digital business models, digital products and services, online communities and social networks, business value of IT, digital health, IT outsourcing, IS adoption, information governance innovation, IS implementation, IT governance and IT project management (Tarafdar et al., 2022). Theory building in IS would be expected to touch upon areas defined or in related extensions in the future, whereby there is a strong IT artifact with which a socio-technical interaction happens at the level of individuals, organizations or society (Struijk et al., 2022). Whilst theory building endeavours using big data driven methods for IS can have both an inductive or deductive approach (Kar & Dwivedi, 2020; Miranda et al., 2022), without such attempts towards theory building, Big Data driven research would fall within the perils of describing data for answering tactical questions (Grover et al., 2020). In doing so, however, care should be given so that such studies attempt to address causal questions that are strategic and not tactical, and move beyond descriptions of the data through mere big data visualization (Kar & Dwivedi, 2020).

3.1 INDUCTIVE THEORY BUILDING THROUGH BIG DATA METHODS

Most of the Big Data driven studies tend to be exciting because they are inductive, and theorizing is based on empirical evidence uncovered in the large volume of data. It may help to break theoretical saturation or enable scholars to establish new dimensions of the phenomenon under investigation either by uncovering new factors or by establishing new relationships among them. While initiating such studies based on research questions, some elements of data selection need to be defined, with sampling strategies that are adequate for the generalizability of the findings, and for establishing the reliability and validity of the results. Furthermore, large datasets often suffer from noise and veracity challenges, which can be addressed by developing proper sampling strategies, whereby spam, bot activities, and misinformation issues can be addressed (Aswani et al., 2019). Data should be collected twice in the process of these studies. Collection of data in the first stage data will act as the basis for exploratory analysis while collection of fresh data after theorizing and model building will work for confirmatory analysis.

After data is acquired in the first stage, it can be analysed using preliminary approaches which could involve methods like topic modelling, sentiment analysis or image recognition, with the objective to derive factors from the data. This is the pattern surfacing stage, whereby possible factors associated with the phenomenon under investigation may be uncovered. This stage of exploratory data analysis uncovers signals from the trace data that has been generated from natural activities. It would be necessary at this point to revisit the literature iteratively and identify theoretical building blocks, which may be useful to explain the phenomenon under

investigation through a more objective lens. Furthermore, it may happen that new factors emerge, which are not documented in the extant literature yet, which may be even more exciting. These factors may be derived from the core IS discipline or from related ones. This process would help to generate theoretical lexicons and pave the path for theory building. After this stage, hypotheses building, or formulation of propositions are encouraged, which could attempt to explain the phenomenon more objectively. It is important to note that the dependent variable (DV) being observed as the most critical in the phenomenon should be measured objectively, in a manner that is different from how the other constructs are being measured to avoid measurement biases. The stage of pattern surfacing can be an initial exploratory study.

After the pattern surfacing stage, it is important to move towards the theoretical model validation stage. This is where hypotheses testing using statistical measures can be undertaken. The end objective in this stage is to move towards a causal inference about the phenomenon under investigation (Mithas & Krishnan, 2009). At this stage, fresh data collection could be initiated using the same sampling strategy. After the data is analysed using Big Data analytics, secondary indicators which have objective numerical measures could be derived. These could be possibly based on term frequency analysis if it is based on natural language processing. A possible direction to move towards could be to use inferential analysis-based approaches to move towards causality by the adoption of heterogeneous treatment and methods. The nature of the data derived after the Big Data analytics would guide the choice of inferential analysis which can be used (Mithas et al., 2022). Robustness checks and sensitivity analysis can enable better trust on the outcomes. Figure 1 showcases these stages through a process diagram.

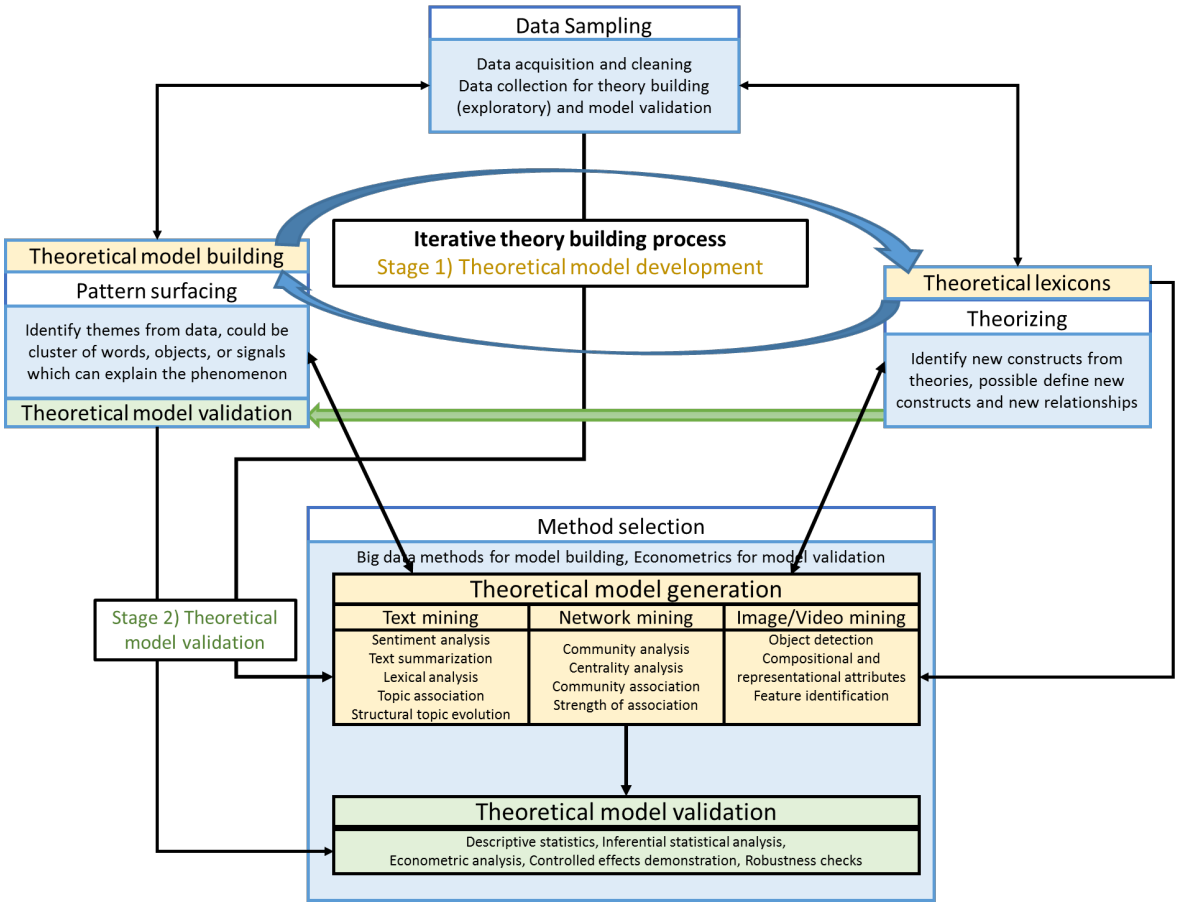

Fig. 1: Process diagram for inductive theory building extending Miranda et al. (2022) and Kar and Dwivedi (2020)

As an example of the above process, in Aswani et al. (2019), pattern surfacing behaviour was determined surrounding social discussions about a particular case study, namely SEOClerks, in the space of digital marketing services. Then after analysing social discussions, patterns were derived which could represent theoretical lexicons from transaction cost economics. These theoretical lexicons were derived through the use of natural language processing and network mining within social media discourses about the platform SEOClerks. Then through a netnography, the study attempted to contribute towards the dark side of digital marketing, by taking the lens of transaction cost economics. The reasons for moving towards a qualitative netnography research was attributed towards studying the negative less represented factors in social discussions in greater depth. However no inferential econometric model building was attempted in this study.

For inductive theory building, Kar (2021) first identified factors from patterns surfacing behaviour, which impact mobile payment use satisfaction by looking at user generated content. Subsequently after factors were mapped to possible theoretical lexicons to build a theoretical model by combining adoption literature and service science literature, the model was validated using inferential data analysis, to establish which factors were statistically relevant. This is important, as Big Data driven studies may generate many factors which may have connect with theoretical lexicons, due to the inherent noise in the data, but not all factors may evolve to form a well-developed theoretical model that can be validated empirically.

Similarly, in Kushwaha et al. (2021) theorizing happened by exploring trace data in social media to identify factors that impact user experiences of chatbots in the exploratory study, and then in the confirmatory study, the factors and their impact on user experience was validated using econometric models. The exploratory study helped in developing theoretical lexicons by identifying the theoretical lens from literature so that there can be reduced scope and better objectivity while looking at the factors while developing the research model. Subsequently the model validation stage established the factors that were actually statistically relevant for impacting the user experience of chatbots by using lasso and ridge regression.

In another example, Kar and Kushwaha (2021) first captured signals through social media surrounding a general theme of users who were discussing about specific artificial intelligence products. Based on identification of theoretical lexicons by mining the text of the social discourses using topic modelling and subsequent network analysis among these topics, hypotheses were developed surrounding factors that impact user experience in these applications. The factors derived were differentiated based on business owners and application users. Subsequently multi-variate data analysis was undertaken to validate the model for improving stakeholder experience for artificial intelligence solutions.

3.2 DEDUCTIVE THEORY BUILDING THROUGH BIG DATA METHODS

For studies involving deductive theory building, the focus of problematization is more towards a gap spotting behaviour within the extant literature (Alvesson & Sandberg, 2011). For deductive studies, the research questions and hypotheses are first identified based on the gaps in the literature; we start, thus, with a strong literature review to identify gaps, with a clear idea of the phenomenon. For deductive theory building, the exploration starts through identifying a gap in the literature. While gap-spotting in social sciences is often a desired strategy for problematization (Tadajewski & Hewer, 2011), the objective would be slightly different than that of other research methodologies in management, which may involve qualitative or quantitative data (i.e., data collected through interviews or surveys). Typically,

the focus should be to establish relationships from real data on context, which may result in biases in primary research (i.e., social desirability biases). In Big Data driven studies, the difference would lie on identifying research questions from gaps, which typically can be addressed by big data analytics of trace data. For example, if user behaviour on digital platforms can have some tangible and measurable changes in the nature of activity, Big Data driven methods can derive factors, which can lead to explaining these changes. Problematization, thus, would depend on insights derivable using Big Data analytics, which may not be measurable objectively through other approaches. This would essentially mean that computationally calculated proxies would be first identified which could represent the theoretical lexicons.

The relationship between such theoretical lexicons can constitute the theoretical model, which would enable mechanisms to abstract findings from the literature. The surfacing patterns, therefore, can lead to the generation of theoretical lexicons, and subsequently give rise to a model that can be abstracted for theory building. The model can be supported by hypotheses that reflect associative or causal relationships between variables. Subsequently, based on research questions and the research model, data sampling strategies may be developed. Data sampling strategies are critical at this stage to ensure two objectives: i) ensure the boundary conditions of the context and subsequently the generalizability of the findings, and ii) reduce the noise in the dataset by carefully identifying systematic methods whereby data collection is undertaken at a context where the phenomenon is actually well represented.

After data collection, the role of Big Data analytics would be to analyse the unstructured data to bring out numerical values based on the unit of analysis (i.e., individual, or firm level). This may be undertaken with computational approaches like natural language processing and image recognition or by using qualitative research methods like content analysis. However, at this stage of analysis, it would be necessary to have mechanisms to deliberate on the reliability and validity of the measures, which are derived to represent the theoretical lexicons.

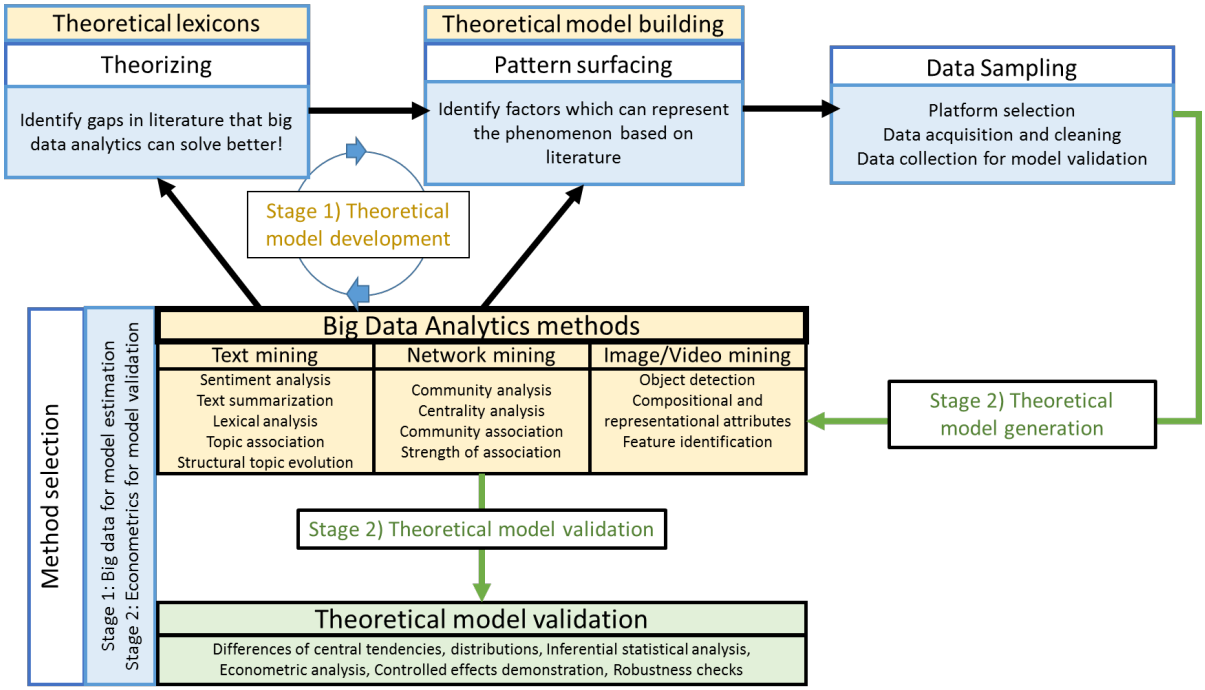

Fig. 2: Process diagram for deductive theory building extending Miranda et al. (2022) and Kar and Dwivedi (2020)

Subsequently, model validation would be undertaken to demonstrate new relationships between the theoretical lexicons in the conceptual model. As with any theoretical model validation, it is possible that some of the hypotheses may not hold true. Similarly, at this stage, during model validation, it would be necessary to develop mechanisms for robustness checks for establishing hypotheses in the model. Similarly, if one is exploring causal relationships, mechanisms to check for reverse causality should be explored. Typically, Big Data driven research may generate strong associative relationships between theoretical lexicons which were otherwise not established, but causal research may need stronger controls, which can typically happen in field experiments. Figure 2 represents the process in a diagrammatic manner.

In one of the early examples of deductive theory building, Oh et al. (2013) demonstrated theory building in the space of misinformation management. During crisis and subsequent discussions in social media, the study focuses on rumour and rumourmongering. First, gaps in literature were identified and hypotheses were developed surrounding anxiety, information ambiguity, personal involvement, and social ties with rumourmongering. Subsequently, data collection was undertaken on Twitter using three case studies, which were globally dispersed geographically, and coding was undertaken to demonstrate both external validity as well as reliability. Then logistic regression was undertaken for the DV and to validate the model, which contributes towards theoretical understanding of rumourmongering.

In similar lines, Wang et al. (2019) undertook a longitudinal case study in a firm for measuring offenses from insider threats and contribute towards theory building in cybersecurity. The theoretical framework was developed first and proxies for measure were identified, which represented constructs like scope of accessed applications, data value of accessed applications, temporal realization, spatial realization, and department size. These constructs were used to predict unauthorized access attempts. Access log data of 51,348 records from 8,588 users from 56 departments was analysed, while the model was validated with a random coefficient model and controlled for fixed effects followed by robustness checks.

Similarly, Grover, Kar, Dwivedi, et al. (2019) explored the phenomenon of political preference polarization through the context of elections and candidate discourse in social media. First, hypotheses were built surrounding differences in voter behaviour in social media. Through the study of social media discourse, reasons behind polarization of voting communities were explained using the model of voter's choice behaviour (Newman & Sheth, 1985). The study derived attributes from twitter discourse using text mining and network mining. These attributes were measured to connect back to theoretical lexicons for explaining polarization behaviour.

As an example of the above process, Gandhi and Kar (2022) studied how Fortune 500 firms build a social presence on social media platforms. The gap of how multi-modal data has rarely been explored for user engagement is first established as an initial problematization. Subsequently, using convoluted neural networks for image mining and natural language processing, the attributes are derived from firm generated content, which formed the independent variables (IV). Using negative binomial regression with log link function, the relationships between the DV, which consisted of different levels of user engagement parameters was established with IV like presence of text in images, text sentiment, message length, presence of face, orientation of face and logo in image posts.

Similarly, Kushwaha et al. (2022) model the role of root level influencers and inherent influencers in changing political preferences and behaviour of social media users. While initial feelers about the context being explored was generated by identifying gaps in the existing literature, subsequently text of the users and the networks of the virtual communities was studies in this article. The study theorizes beyond visualization to explain how the types of

social media influencers may differ in their behaviour and their outcome on their followers. While major findings are data visualization, the inferential analysis attempts to move toward explaining causal incidences among the type of social media influencers and their impacts in the virtual communities.

Other examples of studies which have been published in similar lines can be observed from Grover and Kar (2017) and (Grover, Kar, & Ilavarasan, 2019) where from the mining of big data social signals were converted towards explaining the phenomenon by borrowing theoretical tenets from the information systems discipline.

4. TRAPS TO AVOID IN BIG DATA-DRIVEN THEORY BUILDING

To date, researchers seem to venture into the space of Big Data driven studies, mostly encouraged by the easy access to how data may be collected and the quantum of publications. There are however challenges which emerge in such studies, which we would like to highlight.

Trap 1: Some Big Data studies fall into the trap of picking up an available dataset from an online data hosting platform and analyse it using computational approaches. Alternatively, sometimes the data may be extracted from online platforms either via application programming interfaces or via scraping from websites and analysed with computational approaches. After the analysis, the data is visualized to demonstrate interesting patterns. Such visualization may suffer from challenges of different interpretation to different readers and are rarely falsifiable, and also suffer from reliability and validity challenges in the interpretation. For example, if the data is a large corpus of text, analysis using text mining algorithms and subsequent data visualization may result in interesting graphs and visuals, but they continue to merely describe the data, not necessarily explaining the reasons for a phenomenon. It is important to note, therefore, that data and its description is not a strong theoretical contribution and might not be generalizable beyond a very specific context, or a very specific time period of data collection. However, if the analysis can be abstracted to explain something unique about the phenomenon, then it may constitute the basis for a theoretical contribution.

Trap 2: Some Big Data studies may fall into the trap of not focusing on theory building adequately to explain a phenomenon in a way that is grounded to the boundaries of the nomological network of the theory and extend it. Sometimes, even if a theoretical lens is adopted, there is a lack of theoretical contribution and studies merely demonstrate an application of the theory. Big Data driven studies need to move beyond a cursory treatment of theory, where it becomes just an application of a theoretical lens to actually extend the theoretical boundaries by examining new factors, constructs, and their relationships. The focus should be to move towards explaining causality, with theory being an enabler for explaining the phenomenon in a generalizable way that will stand the test of time.

Trap 3: Some Big Data studies fail to generate theoretical knowledge that would typically be contextualized to the design, use, and impact of the IT artefact under study, and demonstrate a lack of connection with the uniqueness of the IT artefact (Struijk et al., 2022). If one substitutes the IT artefact with another IT artefact, the findings may still remain very similar. This is typically a problem of cursory treatment of the uniqueness of the IT artefact. For example, if a study explores a mobile medicine delivery application, the mobile medicinal delivery application should have elements that are distinct from other mobile applications like a food delivery application. The phenomenon being explored should have distinctive capabilities that differ in both contexts, while majorly both represent a mobile application in their essence.

Trap 4: Finally, some Big Data studies focus on convenience arising from multiple causes. The first cause is access to a particular dataset for data analysis. Data should be collected based on research questions, driven by the phenomenon and research questions should not be generated based on access to data. The second cause is competency in computational methods, which guides data analysis. Since the possibilities of using various data analysis methods is very large, the comfort of using what is known to the team precedes what may be more appropriate for the analysis. The third cause arises from the use of theory as a tool to justify the scope of the work. The research questions should naturally have a fit with the theoretical building blocks, and retrospective fitting of theory based on familiarity is best to be avoided.

5. ARTICLES IN THE SPECIAL ISSUE

The special issue attracted 41 new submissions, across 115 authors, who were spread across 47 institutions across countries including New Zealand, United States, Korea, India, France, Poland, China, Malaysia, Australia, Brazil, Hong Kong, and Iran. Four out of these submissions were finally accepted in our special issue. King and Wang (2023) address the timely topic of misinformation diffusion and examine the spread of authentic news and misinformation as well as the amplification of real *versus* misinformation during crisis situations, using big data to validate such relationships. The authors have used the deductive approach to validate the model through hypothesis testing, showing that virality is higher for misinformation, novel tweets, and tweets with negative sentiment or lower lexical density. Subsequently, (Joung & Kim, 2023) have performed customer segmentation using a machine-learning approach for product development based on the importance of product features from online product reviews. Using an inductive method, they identify and interpret the nonlinear relations between satisfaction with product features and overall customer satisfaction. Similarly, Zhang et al. (2023) examine how alternative food networks cultivate engagement on a social media platform. They use an inductive approach and have collected data from social media, and it is being validated empirically through an LDA model. Their empirical results demonstrate that posts centred on openness/disclosure, sharing of tasks, and knowledge sharing result in positive levels of social media engagement. Finally, Almaqableh et al. (2023) undertake an event study to investigates the link between cryptocurrency markets and drug trafficking activities. Their study confirms the predictions of convenience theories of crime as to the relative attractiveness of cryptocurrencies to criminals, and the extent to which not only general, but also their own future interests, sacrificed readily on the altar of accessibility.

7 | CONCLUSION

In this editorial, we extend the directions of Kar and Dwivedi (2020) and Miranda et al. (2022) to provide further guidance in Big Data driven research, especially for publishing in established information systems journals. In particular, we focus on extending the discussions by demarcating between inductive theory building and deductive theory building in Big Data driven research. We establish directions on the processes that researchers may undertake in different stages, so that both theory generation and theory validation may be attempted in these lines of enquiry. We also highlight common pitfalls of such studies and warn future researchers to avoid them, so that the studies contribute better towards inferential analysis in IS research.

REFERENCES

Almaqableh, L., Wallace, D., Pereira, V., Ramiah, V., Wood, G., Veron, J. F., Moosa, I., & Watson, A. (2023). Is it possible to establish the link between drug busts and the cryptocurrency market? Yes, we can. *International Journal of Information Management*, 102488.

Alvesson, M., & Sandberg, J. (2011). Generating research questions through problematization. *Academy of management review*, *36*(2), 247-271.

Angelopoulos, S., Bendoly, E., Fransoo, J. C., Hoberg, K., Ou, C. X., & Tenhiälä, A. (2023). Digital transformation in operations management: Fundamental change through agency reversal. *Journal of operations management*, *Forthcoming*.

Angelopoulos, S., Brown, M., McAuley, D., Merali, Y., Mortier, R., & Price, D. (2021). Stewardship of personal data on social networking sites. *International Journal of Information Management*, *56*, 102208.

Aswani, R., Kar, A. K., & Ilavarasan, P. V. (2019). Experience: managing misinformation in social media—insights for policymakers from Twitter analytics. *Journal of Data and Information Quality (JDIQ)*, *12*(1), 1-18.

Barki, H., Rivard, S., & Talbot, J. (1993). A keyword classification scheme for IS research literature: an update. *MIS quarterly*, 209-226.

Benbasat, I., & Zmud, R. W. (2003). The identity crisis within the IS discipline: Defining and communicating the discipline's core properties. *MIS quarterly*, 183-194.

boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165-1188.

Davison, R. M. (2022). Impact and implications for practise. *Information Systems Journal*, 1-5.

Gandhi, M., & Kar, A. K. (2022). How do Fortune firms build a social presence on social media platforms? Insights from multi-modal analytics. *Technological Forecasting and Social Change*, *182*, 121829.

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. *International Journal of Information Management*, *51*, 102048.

Grover, P., & Kar, A. K. (2017). Big data analytics: A review on theoretical contributions and tools used in literature. *Global Journal of Flexible Systems Management*, *18*, 203-229.

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes–Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, *145*, 438-460.

Grover, P., Kar, A. K., & Ilavarasan, P. V. (2019). Impact of corporate social responsibility on reputation—Insights from tweets on sustainable development goals by CEOs. *International Journal of Information Management*, *48*, 39-52.

Grover, V., Lindberg, A., Benbasat, I., & Lyytinen, K. (2020). The perils and promises of big data research in information systems. *Journal of the Association for Information Systems*, *21*(2), 9.

Heising, L., & Angelopoulos, S. (2021). Early diagnosis of mild cognitive impairment with 2-dimensional convolutional neural network classification of magnetic resonance images.

Heising, L., & Angelopoulos, S. (2022). Operationalising fairness in medical AI adoption: detection of early Alzheimer's disease with 2D CNN. *BMJ Health & Care Informatics*, *29*(1).

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, *70*, 102641.

Kar, A. K. (2021). What affects usage satisfaction in mobile payments? Modelling user generated content to develop the "digital service usage satisfaction model". *Information Systems Frontiers*, *23*, 1341-1361.

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research–Moving away from the "What" towards the "Why". *International Journal of Information Management*, *54*, 102205.

Kar, A. K., & Kushwaha, A. K. (2021). Facilitators and barriers of artificial intelligence adoption in business–insights from opinions using big data analytics. *Information Systems Frontiers*, 1-24.

Kerr, D. S. (1975). Proceedings of the International Conference on Very Large Data Bases. ACM, Framingham, Massachusetts, USA.

King, K. K., & Wang, B. (2023). Diffusion of real versus misinformation during a crisis event: A big data-driven approach. *International Journal of Information Management*, 102390.

Kushwaha, A. K., Kar, A. K., Roy, S. K., & Ilavarasan, P. V. (2022). Capricious opinions: A study of polarization of social media groups. *Government Information Quarterly*, *39*(3), 101709.

Kushwaha, A. K., Kumar, P., & Kar, A. K. (2021). What impacts customer experience for B2B enterprises on using AI-enabled chatbots? Insights from Big data analytics. *Industrial Marketing Management*, *98*, 207-221.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., & Gutmann, M. (2009). Computational social science. *science*, *323*(5915), 721-723.

Malakar, P., George, T., Kumar, S., Mittal, R., Natarajan, V., Sabharwal, Y., Saxena, V., & Vadhiyar, S. S. (2013). A divide and conquer strategy for scaling weather simulations with multiple regions of interest. *Scientific Programming*, *21*(3-4), 93-107.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, *2*(1), 460-475.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Marx, V. (2013). The big challenges of big data. *nature*, *498*(7453), 255-260.

Merv, A. (2011). Big Data: It's going mainstream, and it's your next opportunity. *Teradata Magazine*, *1:11*. https://web.archive.org/web/20110218053426/www.teradatamagazine.com/v11n01/Features/Big-Data/

Miranda, S., Berente, N., Seidel, S., Safadi, H., & Burton-Jones, A. (2022). Editor's Comments: Computationally Intensive Theory Construction: A Primer for Authors and Reviewers. *MIS quarterly*, *46*(2), iii-xviii.

Mithas, S., & Krishnan, M. S. (2009). From association to causation via a potential outcomes approach. *Information systems research*, *20*(2), 295-313.

Mithas, S., Xue, L., Huang, N., & Burton-Jones, A. (2022). Editor's Comments: Causality Meets Diversity in Information Systems Research. *MIS quarterly*, *46*(3), iii-xviii.

Newman, B. I., & Sheth, J. N. (1985). A model of primary voter behavior. *Journal of Consumer Research*, *12*(2), 178-187.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS quarterly*, 407-426.

Price, D., McAuley, D., Mortier, R., Greenhalgh, C., Brown, M., & Angelopoulos, S. (2015). Inter-social-networking: Accounting for multiple identities. Social Computing and Social Media: 7th International Conference, SCSM 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings 7,

Sidorova, A., Evangelopoulos, N., Valacich, J. S., & Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. *MIS quarterly*, 467-482.

Struijk, M., Angelopoulos, S., Ou, C., & Davison, R. M. (2020). Influencing information quality: Evidence from a military organization. European Conference on Information Systems,

Struijk, M., Angelopoulos, S., Ou, C. X., & Davison, R. M. (2023). Navigating Digital Transformation Through an Information Quality Strategy: Evidence From a Military Organization. *Information Systems Journal*, *33*(4).

Struijk, M., Ou, C. X., Davison, R. M., & Angelopoulos, S. (2022). Putting the is back into is research. *Information Systems Journal*, *32*(3), 1-4.

Tadajewski, M., & Hewer, P. (2011). Intellectual contributions and 'gap-spotting'. In (Vol. 27, pp. 449-457): Taylor & Francis.

Tarafdar, M., Shan, G., Bennett Thatcher, J., & Gupta, A. (2022). Intellectual Diversity in IS Research: Discipline-Based Conceptualization and an Illustration from Information Systems Research. *Information systems research*.

Venters, W., Oborn, E., & Barrett, M. (2014). A trichordal temporal approach to digital coordination. *MIS quarterly*, *38*(3), 927-A918.

Wang, J., Shan, Z., Gupta, M., & Rao, H. R. (2019). A longitudinal study of unauthorized access attempts on information systems: The role of opportunity contexts. *MIS quarterly*, *43*(2), 601-622.

Williams, D. (2012). *If 'Big Data' Simply Meant Lots of Data, We Would Call It 'Lots of Data'*. Forbes.

Zhang, Y., Ridings, C., & Semenov, A. (2023). What to post? Understanding engagement cultivation in microblogging with big data-driven theory building. *International Journal of Information Management*, 102509.