Theory Building with Big Data-Driven Research – Moving away from the "What" towards the "Why"

Arpan Kumar Kar
Department of Management Studies, Indian Institute of Technology Delhi, India

Yogesh K. Dwivedi
Emerging Markets Research Centre (EMaRC), School of Management, Swansea University, UK

Abstract: Data availability and access to various platforms, is changing the nature of Information Systems (IS) studies. Such studies often use large datasets, which may incorporate structured and unstructured data, from various platforms. The questions that such papers address, in turn, may attempt to use methods from computational science like sentiment mining, text mining, network science and image analytics to derive insights. However, there is often a weak theoretical contribution in many of these studies. We point out the need for such studies to contribute back to the IS discipline, whereby findings can explain more about the phenomenon surrounding the interaction of people with technology artefacts and the ecosystem within which these contextual usage is situated. Our opinion paper attempts to address this gap and provide insights on the methodological adaptations required in "big data studies" to be converted into "IS research" and contribute to theory building in information systems.

Keywords: Big data analytics; Image mining; Network mining; Sentiment analysis; Text mining; Inductive theory building; Machine learning; Information management.

1. Introduction

The availability and access to data has changed, as across organizations and nations, digital transformation initiatives are maturing. Increasingly, people get ownership of digital devices and access to the internet. As a result, footprints are created with similar pace, when people connect with each other to share information, consumer experiences, concerns and other user generated content (UGC) online, especially in social media and social commerce platforms. Electronic markets like Amazon, eBay, and Rakuten allow consumers to share their experiences through feedback and ratings. Similarly, the sharing economy provides platforms where service providers and consumers connect, and the feedback empower citizens. Concurrently, platforms like MyGov in India and Singapore, allow citizens to give feedback on government policies and focus on citizen empowerment, which also contributes UGC. Sensors in smart and wearable devices also contribute to the generation of big data. Robotics, RFID and Industry 4.0 technologies like sensors and actuators also contribute to the growth of big data information assets that organizations use for value creation (Curtin et al., 2007; Chakraborty and Gupta, 2016; Chakraborty and Joe, 2017; Fosso Wamba et al., 2017). Such endeavours create huge repositories of UGC, enabling researchers to easily access huge amounts of data. Further computational technologies like cloud computing and super-computing platforms are facilitating in creating an environment where this data may be analysed easily (Kar, 2017). Therefore, big data may be collected by researchers from online forums, social media, emails, news and online forums, devices connected through the internet of things (IoT), telecommunication devices, sensor-based applications in devices, and even from more than one source in multi-modal studies. Security and information risk management is also critical for such big data (Singhal and Kar, 2015). Big data is often used for research and studies are being developed based on analysis of such resources. This big data is characterized by both structured and unstructured data having high volume, high velocity, high variety and high veracity (Grover and Kar, 2017).

Before the escalation in the volume and granularity of big data, scholars pursuing research in Information Systems (IS) used to face major challenges when it came to data access (Lyytinen, 2009). Enterprise-level data from internal organizational IS, were a valuable source, though not many researchers could get access. The problem of access to data, however, is often reduced for many researchers in this domain. Due to the increasing focus towards digital transformation and growth of internet-based platforms, data collection through APIs or open-access datasets became easier (or sometimes through web-scraping). Further, this has led to the growth of open-data projects whereby organizations and governments collaborate to openly share data to create innovative models of services to empower consumers. This data availability has created a focus towards data-driven research in IS due to the data volume, ease of demonstrating statistical significance, objectivity to storytelling, and availability of computational tools (Grover et al., 2020). However, this data-driven IS research demonstratively still lacks a connection with the theoretical building blocks, which come from management theory, organization theory, behavioural theory, computer science theories, and systems theory (Barki et al., 1993; Dwivedi et al., 2011). Apart from the core computer science theories, the other disciplines enable IS scholars to theorize how consumers interact with technology within individual, organizational, social, and political contexts, and capture the impact or economic implications of such interaction. These studies using big data-driven methods may focus on developing a better understanding of applications in domains like e-commerce and market intelligence, e-government and politics, innovation in science and technology, smart health and well being, security and public safety (Bharathi, 2017; Chen et al., 2012). However, the increased access to and use of big data has disrupted the development of theory in information management, whereby the essence of contributing to these neighbouring disciplines is increasingly lacking, be it at the individual, organizational, social, and political levels of analysis.

Further, it was highlighted by Pentland (2008) how the concept of honest signals and its interpretation is typically enhanced in big data-driven studies. The analysis of big data can really offer a view on psychological and behavioral elements that could not be observed before. Moreover, big data driven methods allow researchers to undertake a view on spontaneous behaviors that are difficult to change even when subjects know of being observed (Fronzetti Colladon, 2018). Further, this also helps in undertaking research when the researchers are distanced from the subject when the context of examination happens (Kar, 2020). This approach using big data analytics therefore solves some of the biases a study could have while administering a survey-based research methodology. Similarly, big data-driven studies by mining these honest signals can reveal unconscious behaviours and emotional states of consumers and users before they become conscious. For example, big data driven methods have been used to anticipate employees' disengagement (Gloor, Fronzetti Colladon, Grippa, et al., 2017b) by looking at unconscious changes in communication and knowledge exchange behaviours. This makes an exploration possible to understand a mirror of virtual behaviors, ultimately affecting the self-awareness and performance of employees (Gloor et al., 2017a).

IS research has spawned from computer science and demonstrates distinct characteristics of its own. Theory development in computer science (in the space of algorithms) takes place through theoretical validation of complex mathematical phenomena or experimentation, to address improvement of algorithmic capabilities in terms of accuracies of prediction, improved computational complexity or improved time complexity. Such computational contributions can use data readily available in the public domain for experimentation and subsequent theory building. When it comes to IS research, however, we do not restrict the focus on development of new algorithms and validate their strengths or weaknesses. Our focus has increasingly turned towards data collection and analysis and how these activities reflect findings. Therefore, the connection with theory building is gradually getting lost, wherein studies should contribute to management theory, organization theory, behavioural theory, social science theory and systems theory based on analysis of information available from IS. Most of recent big data studies tend to focus on "what is hidden inside the data" rather than attempting to explain

"why this is so". Thus these studies may suffer from limitations as the models may have problems of overfitting based on data, and not accurately predict the future and other similar contexts, making them less generalizable. Further these studies cannot generate knowledge about the relationships among factors examined within the context (Grover, 2020). Thus there is a need for the intersection of data-driven research and theory-driven research during these era of digital transformation whereby not only big data is accessible but also the tools to analyse them efficiently (Maass et al., 2018).

In this context, we focus on providing directions towards the following challenges faced by researchers working on big data problems in IS discipline, with a specific focus on theory building based on UGC:

1. How can authors move beyond presenting "what the data captures" and move towards "why it is so" or "how can the factors in the visualization" be validated objectively?
2. How can authors attempt to focus on in their research methodology to introduce greater objectivity in the outcome through rigor in the methodology?

In doing so, we provide a brief description on what constitutes theoretical contribution in the era of big data and discuss the ways in which IS researchers can contribute to theory building in the discipline. The rest of the article covers first a brief overview of what constitutes theory building in IS discipline, then approaches for theory building and validation, then how to build trust on findings of big data, followed by a discussion and conclusion.

2. Revisiting theoretical contributions for data driven research

Theory development within the IS discipline is likely to conform to expectations of theory in management overall, with some extensions. However, theorizing in management itself has also witnessed a lot of debate as to what theorizing is and is not (Sutton and Staw, 1995; Weick, 1995). In general, however, there is consensus that describing the data in itself is not theorizing but can assist in theory development whereby the researchers would prescribe generalizable models, which could be validated by data to explain a complex phenomenon (Whetten, 1989). IS research and theories are no different, apart from having a stronger relationship to technology artifacts (Dwivedi et al., 2011). Theoretical discussions in IS have typically attempted to address domain, socio-political, structural or ontological, and epistemological questions (Gregor, 2006). We envision that with research using big data analytics, IS researchers could typically attempt to address domain questions and socio-political questions more as compared to the other questions. A good theoretical model should attempt to explore a phenomenon that involves the use of or interaction with Information and Communication Technologies (ICTs) artifacts, whereby it is able to explain and predict some elements of the behaviour in a way that is somewhat objectively testable. Certain concepts or constructs in this context should be identifiable, their relationships should be specifiable, and their relationship should be testable and hence falsifiable. Hence, IS research is often centered around measurement of otherwise abstract concepts so that complex relationships in the theoretical model be specifiable and testable (Subramanian and Nilakanta, 1994). This is a more suitable definition in the era of data-driven research, whereby we move away from the interpretive paradigms to a more positivist paradigm of research. This interaction can be within and across entities like individuals, organizations, society and polity. The phenomenon could be related to how these entities use, impact or are impacted by the technology artifacts.

While prior IS studies conducted were governed by both positivist and interpretive paradigms, data-driven research based on big data needs to be more focused towards the positivist paradigm, whereby using statistical measures and inferential analysis, propositions and hypotheses need to be validated. This is because the data captures a lot of veracity, which otherwise can cloud the lens undertaken by

the researchers. Findings from natural language processing, social network analysis, or other big data analytics approaches typically present findings using visualization-inferences, which merely describe the data, and are atheoretical in nature. However, such approaches may help to uncover factors or constructs which may be further used for theoretical model development. Researchers need to report beyond "what" is hidden in the data to "how" such factors are inter-related and "why" they behave in a particular manner (Whetten, 1989). In such an attempt, it is necessary to constitute a disciplined imagination to identify suitable factors and explore relationships between them so as to balance the trade-off between parsimony and comprehensiveness (Weick, 1989; Whetten, 1989), since such methods can help to discover too many elements in the data which may create a chaotic interpretation. Since the readers and reviewers are often distanced from the actual phenomenon and data, researchers' interpretations are otherwise difficult to be communicated and absorbed without dissonance, unless greater objectivity can be introduced in the research methodology.

Therefore, IS researchers using big data analytics need to ensure that the connection with theory lenses from the existing core IS as well as neighbouring management disciplines are strongly developed while looking at a phenomenon that involves interaction of the entities (e.g. individuals, organizations, society, and policy) with technology artifacts (e.g. social media, smart technologies, IoT, wearables, mobile devices, digital services and other data generating platforms). For example, these studies may attempt to view technology usage or adoption through a particular lens for a particular stakeholder like individuals, and then a suitable framework may be adopted like UTAUT or UMEGA depending on the stakeholder (Dwivedi et al., 2019; Dwivedi et al., 2017). Further, when such technologies are utilized within organizations, organizational theories may be more relevant (Kar, 2015; Business-Frontiers, 2020). Findings from these studies would be expected to be sufficiently abstracted so that knowledge may be grounded back to existing literature through the development of management theories, organization theories, behavioural theories, and systems theories. Research methodologies would also have to be appropriately developed so that findings from the data have greater trust among the audience. A brief overview of this journey is presented in figure 1.

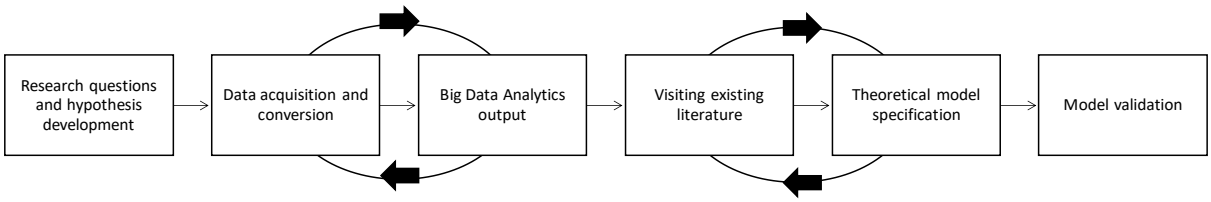

Figure 1: Journey towards theory development in big data research

The first stage of this journey should begin with the broad research questions and identification of the theoretical lens. Subsequently such a journey would start with data acquisition and conversion based on research questions, the former would define the data acquisition and sampling strategy. Then the data needs to be analysed with approaches and methods of big data analytics. Then the researchers need to revisit the existing literature to understand the findings and develop a theoretical model surrounding the research questions and hypothesis. Multiple iterations may be required in the second, third, and fourth stages of the inductive approach towards theoretical model building and subsequently model validation. Model specification based on the output of big data analytics at this stage needs to be validated using statistical, and inferential methodologies. We elaborate in the following section, where we attempt to identify approaches to meet desired research objectives.

3. Approaches for theory building and validation

By specifying possible research methodologies, we develop guidelines towards theory building for addressing grand problems at individual, organizational, social or political levels. So at the first stage, identifying interesting theoretical questions which would have a larger interest and allow validation using the data, needs to be specified. In particular, for such big data studies, data would demonstrate elements of high volume, velocity, variety, variability, veracity, visualization, and leading to value in understanding the focal phenomenon. These studies could use data available from open or primary data sources, which could be behavioural or usage-based from the relevant stakeholders. Since big data has a lot of "noise" due to the high variety, variability, and veracity, bringing objectivity systematically is critical for the scientific sanctity of findings. Applications of artificial intelligence like machine learning and deep neural network algorithms would be key in this context, and so a detailed understanding of the computational scope of the background algorithms is also required so that the outcomes of studies are more reliable and generalizable.

This data would be basically driven by the presence of platforms which create and allow usage of data like social media, application, sensor and location-based data (Lee at al., 2019; Misirlis and Vlachopoulou, 2018; Rathore et al., 2017). For example, by using social media APIs, it is possible to extract posts based on hashtag or keyword-based search in a platform like Twitter. Similarly, it is possible to extract the tweets of the timeline of an individual user, if the tweets are public. In the first case, the unit of analysis is the content of the tweet on a specific context. In the second case, the unit of analysis could be the user or consumer. Similarly, it may be possible to access mobility data of individuals using mobile crowdsensing applications or social media data (Wagner, 2020). Further it may be possible to analyses both individual (user or consumer) behaviour and group behaviour based on data from sensors (Chaffin et al., 2017). Further, based on networks of mobility, it may be possible to predict human behaviour connected to a context that drives this mobility, for theorizing in a context like pandemic management. It is important to deconstruct the mobility variables beyond the descriptive elements so that an inferential theoretical model may be built later. Thus, it is important to frame research questions with sufficient clarity so that one can trace the unit of analysis and what the research model attempts to validate.

In particular, studies need to demonstrate how the researcher's systematic biases were attempted to be reduced in such methodologies at every stage. The challenge starts with data acquisition, and the need to minimize usage of data collected for other purposes and reuse them by force-fitting them to problem statements. While computational contribution may be feasible on such data, such contribution to the IS theory may be extremely limited with such an approach. That being said, not all primary datasets may be useful unless the collection is conducted appropriately (Tirunillai and Tellis, 2012; George et al., 2016). For example, if tweets are extracted based on keywords, a single keyword may not provide ample result if the theme for discussion is not very popular (Grover et al., 2020; Kushwaha et al., 2020). How does one identify the rest of the keywords objectively which may help to get tweets in similar themes? Developing a keyword association matrix and identifying the top 10 keywords from a smaller subset of few thousand tweets, may solve this problem of objectively identifying associated keywords (Joseph et al., 2017; Georgiadou et al., 2020; Grover et al., 2020). After collection of large volumes of user generated content, it may be necessary to remove a large part of data which do not meaningfully contribute to the research question. In many cases, retweets and tweets which do not have enough content for analysis to be meaningful, could be eliminated. Sampling within the dataset is extremely critical and needs to be undertaken judiciously to minimize veracity wherever possible (Grover et al., 2020; Mishra and Singh, 2018). These approaches can minimize trade-offs between internal and external validity problems and bring greater reproducibility of findings. Thus, after extraction, it is important to undertake data cleaning based on research questions, through approaches like stemming (Paice, 1994).

Many studies at present restrict the focus on collecting data, and with visual inferences demonstrate what the data captures in essence in such data-driven research. For example, using sentiment analysis it is possible to check the polarity or sentiment of UGC, map this into positive, negative or neutral posts or topics, and connect it to a context for theory building (Kar, 2020). Such a context could be related to an issue related to service usage or related to an event like Brexit (Aswani et al, 2018; Georgiadou et al., 2020),. Similarly, using other approaches of text mining like topic modelling (Wallach, 2006) and opinion summarization (Wu et al., 2020), large volumes of text can be summarized into closely grouped themes. Approaches like *latent dirichlet allocation*, *word2vec* and *n-grams* also help in such text summarization-based studies, especially in microblogging data (Hannigan et al., 2019). Studies often represent these output of text summarization using word clouds, whereby words across topics which have larger occurrences have larger size in the visualization and enables visual inferences to be developed (Cui et al, 2010). Similarly, text mining and natural language processing can help to derive attributes that may be modelled by referring back to existing literature to answer the research questions (Tirunillai and Tellis, 2012; Dong et al, 2018). Social network analysis enables studies to be conducted whereby visual graphs based on cooccurrence of words or keywords can be developed (Barabasi, 2016). Using such parameters, it may be possible to develop theories surrounding network level attributes like the nature of connections between the entities, the closeness of their association (cliquishness), the strength of ties, the structural holes or bridges, the nature of directionality of interaction (Zuo et al., 2020). Further it may be possible to theorize based on user level attributes like homophily, propinquity, reciprocity and multiplexity (Barabasi, 2016). Through bibliographic analysis, network-based approaches enable the development of word association and co-author association networks in many literature reviews. Content analysis approaches has also been attempted in conjunction with network analysis of unstructured data (Wu, 2013; Angelopoulos and Merali, 2017). Also, theoretical models may explore images in conjunction with text and networks. Image characteristics like image content-based attributes like direct images, linked images have been used. Further image characteristics like colourfulness, presence of human faces, source of image, image quality, image text fit has been used (Li and Xie, 2019).

A lot of times however, studies do not provide theoretical contribution in specific areas, and their research questions are restricted to propositions validated through visual inferences, which might not be problematic for an exploratory stage (Chae, 2015). However, for mature contexts, this approach may negate the objective to build theory. This is a problem that studies should try to address while attempting theory building in IS using big data analytics, as most studies attempts to visually represent data without explaining why and how factors that are identified from the data behave, as highlighted in editorial reviews (Grover, 2020; Grover et al, 2020). So what? These studies may lack a strongly defined dependent variable, an explanation as to why findings are so, how visual elements relate to each other, and how they explain the focal phenomenon objectively through model specification.

Therefore, for hypothesis validation, at the first step, if visually one undertakes analysis of large volumes of text using text summarization, one needs to take a step further from mere visual inferences. Sometimes in such contexts methodologies like field experiments and content analysis methods for consumer research can helpful for building reliability and validity in the research model (Kassarijan, 1977; Lambrecht and Tucker; 2013). In content analysis, a word can be a unit of analysis which can be mapped to a theme or a construct in a Likert-like scale. Similarly, topic models derived after text summarization can be mapped to constructs or themes of a research model. Content analysis may be automated by building a pool of words, which could represent common themes, or constructs. It would be necessary to describe the construct well and establish its face validity in the focal context. Face validity of the concepts and how they are connected together could be done through a focused group interview or through clustering of words based on cooccurrences (Kar, 2020). Alternately, existing

theories could be useful for establishing validity in these contexts. These methods could also be attempted for naming clusters of keywords after developing a network of keywords based on cooccurrences in the topic models. This will also help to identify and map the words to specific constructs from existing literature, and may even enable the development of a new construct for a proposed model. It would be also important to build mechanisms of *category reliability* and *inter-coder (judge) reliability* within this approach. This would require human intervention by a team of researchers who are familiar with the domain of research well. Since there is too much veracity and volume in the data, the analysis using big data methods can create outputs which require an interdisciplinary lens to be validated objectively. The studies should always report on the outputs generated from such validity and reliability tests. Alternatively, methods like automated content analysis based on term frequency and bag of words can be used in such studies whereby *natural language processing* is used extensively for establishing reliability and validity of context mapping (Grover and Kar, 2020; Kang et al., 2020). Such approaches for establishing reliability and validity by analysing unstructured data have been demonstrated by Saxton et al. (2019) for examining specifics of CSR related tweets. Similarly, Oh et al. (2015) demonstrates how content analysis methodology borrowed from social science research can be applied to social media analytics focused research based on collective sense making. Measurement is integral to IS research and the era of data-driven research is no different. So, measurement of concepts and model specification becomes important parameters to trust the outcome of model validation (Subramanian and Nilakanta, 1994).

At the second stage of actual validation of hypothesis, it would be advisable to bring out statistical measures. After analysis is conducted, it may be feasible to identify and map content to one or more groups of constructs or themes. Such an approach could attempt to validate the statistical significance of the number of instances of occurrences in each group. The distribution of the occurrences could be validated with statistical methods which test for the differences of distribution in terms of mean values or dispersion (Grover et al., 2020). Statistical tests are important to understand the degree of overlap between the possible groups and can provide better reliability of findings. Sometimes, simple statistical tests like *T-tests*, *Z tests* and *ANOVA* could also be used to statistically validate such differences (Grover et al, 2019b). However such statistical validation would not answer why some observations are occurring still but would enable a more objective evaluation of the similarity or differences in outcome of observations within the groups of big data. These would however be intermediate validation and scope for more analysis would emerge.

The final objective of such studies should be building inferential research models. A good understanding of advanced econometrics and machine learning would be helpful in such a context. Because of the veracity of big data, it may not be possible to always attempt advanced multivariate analysis and authors may be forced to produce descriptive statistics only (Aswani et al., 2018; Mishra and Singh, 2018). However, approaches like principal component analysis and penalized regression may help in dimensionality reduction and variable selection in such cases (George et al., 2016). The reason for this methodological limitation may arise because of the actual number of observations (rows) which can be used to validate a particular model may not be very high. For example, a million tweets or reviews on a specific theme (represented by a hashtag) may have only a hundred representative topics (Hannigan et al., 2019). These latent topics may be labelled using *entropy-based approaches* and mapped with literature (Tirunillai and Tellis, 2012). Now these hundred topics may be mapped to over a dozen concepts using content analysis methods as elaborated earlier. However, since each topic corresponds to one single instance, hundred topics would actually correspond to a sample size of hundred, for a model with a large number of constructs. Such a model may be difficult to validate using traditional approaches like multi-variate analysis because of noise in the data (Gefen et al., 2000). However dimensionality reduction may address this limitation using approaches like *Bayesian*

regression and *regression trees* (George et al., 2016). However, attempts should always be made towards validating a non-parametric econometric model if possible where the dependent variable is captured objectively and computationally. If not possible, at the bare minimum, attempts can be made towards validating a multiple regression model where there are independent variables which are used to predict one or more dependent variables, where in control variables are also identified and introduced in the models (Kar, 2020). Advanced econometric models that can be used in such big data research include *ridge and lasso regression, principal components regression, partial least squares, bayesian variable selection, and regression trees* (Varian, 2014).

Further advanced machine learning algorithms like deep learning and convoluted neural networks may also be interesting in such cases (LeCun et al., 2015; Schmidhuber, 2015; Gu et al., 2018). Machine learning methods and algorithms like *deep learning, neural networks, computer vision, image classification, support vector machine, genetic algorithm, ant colony optimization, porter stemmer algorithm, conditional random field algorithm, penalised regression, viola-jones algorithm, quantile regression, supervised learning, semi-supervised learning and unsupervised learning has been used in marketing and finance literature in recent times* (Ali and Kar; 2018; Balducci and Marinova, 2018; Henrique et al., 2019). Such machine learning algorithms replicate the way biological organisms process information to develop learning, adaptivity and reasoning capabilities while predicting outcome from both structured and unstructured data. Such studies using machine learning may use algorithms like *neural networks, genetic algorithms, swarm intelligence* and other biologically inspired algorithms and swarm intelligence which can operate in the metaheuristics and hyperheuristics state (Chakraborty and Kar, 2016; Kar, 2016; Chakraborty and Kar, 2017). However machine learning and artificial intelligence based algorithms operate as a blackbox and inferential relationships are often difficult to establish from the viewpoint of *replicability, explainability, fairness* and *transparency* (Duan et al., 2019; Dwivedi et al., 2019). Care should be taken not only to establish causality relationships but also explore the possibilities of reverse causality wherever possible. Further it is important to note that model specification based on data description would be critical to report in all such cases, along with model estimation and results (Adamopoulos et al., 2018; Goh et al., 2013; Oh et al., 2015).

Using mixed-research methods like netnography, field experiments or content analysis methods, it may be possible to identify independent variables and score them across instances (George et al., 2016; Saxton et al., 2019). However the dependent variable is always preferably a computed one based on the nature of the research question and the unit of analysis (e.g. Dong et al., 2018; Salehan and Kim, 2018; Singh et al., 2020). A lot of derived metrics are available in existing literature which could be used for such definition of dependent variables. Such a computed variable or a mix of such derived variables could be used as a proxy to measure a construct from existing literature. A mix of such derived variable could be a better proxy measure for validating a theoretical model. Such mix derived attributes could be obtained from a review of literature on social media analytics (Chae, 2015; Ghani et al., 2019; Lee, 2018; He at al., 2019; Rathore et al, 2018; Xiang et al., 2017). Further a mix of variables may be taken from objective and structured data and other variables may be computed from unstructured data (Dong et al., 2018). Given the nature of big data, shifts between reflective measurement models and formative measurement models would happen across studies based on their scope, but establishing objectivity and reporting it would be important.

These stages of theoretical model development would require continuously to relook at existing literature from IS and other related disciplines, to identify possible alternate measure for existing constructs (here referred to as proxies). Some of these relevant computed factors may not have any suitable construct or concept identified in academic literature. This gives the research team an opportunity to bring out a new measurement for a possible factor that can be grounded back to the

theory for enriching it. The elaborated approaches with representative activities or methodologies which could be adopted are explained through the block diagram in figure 2.

The "Big Data Analytics block" in figure 2 helps in theorizing by identifying attributes and factors which otherwise are difficult to measure or estimate. These factors would need to be aligned with a theoretical lens iteratively for the process of theorizing. After theorizing, using inferential analysis and statistical methods, the theoretical model would be required to be validated. Theory building and theory validation are two different elements which require utmost focus in these studies, since unless a model is validated, theorizing may not be very helpful. This challenge would arise because big data would typically have too much of noise which would often not allow an objective assessment of the outputs of the "Big Data Analytics block".

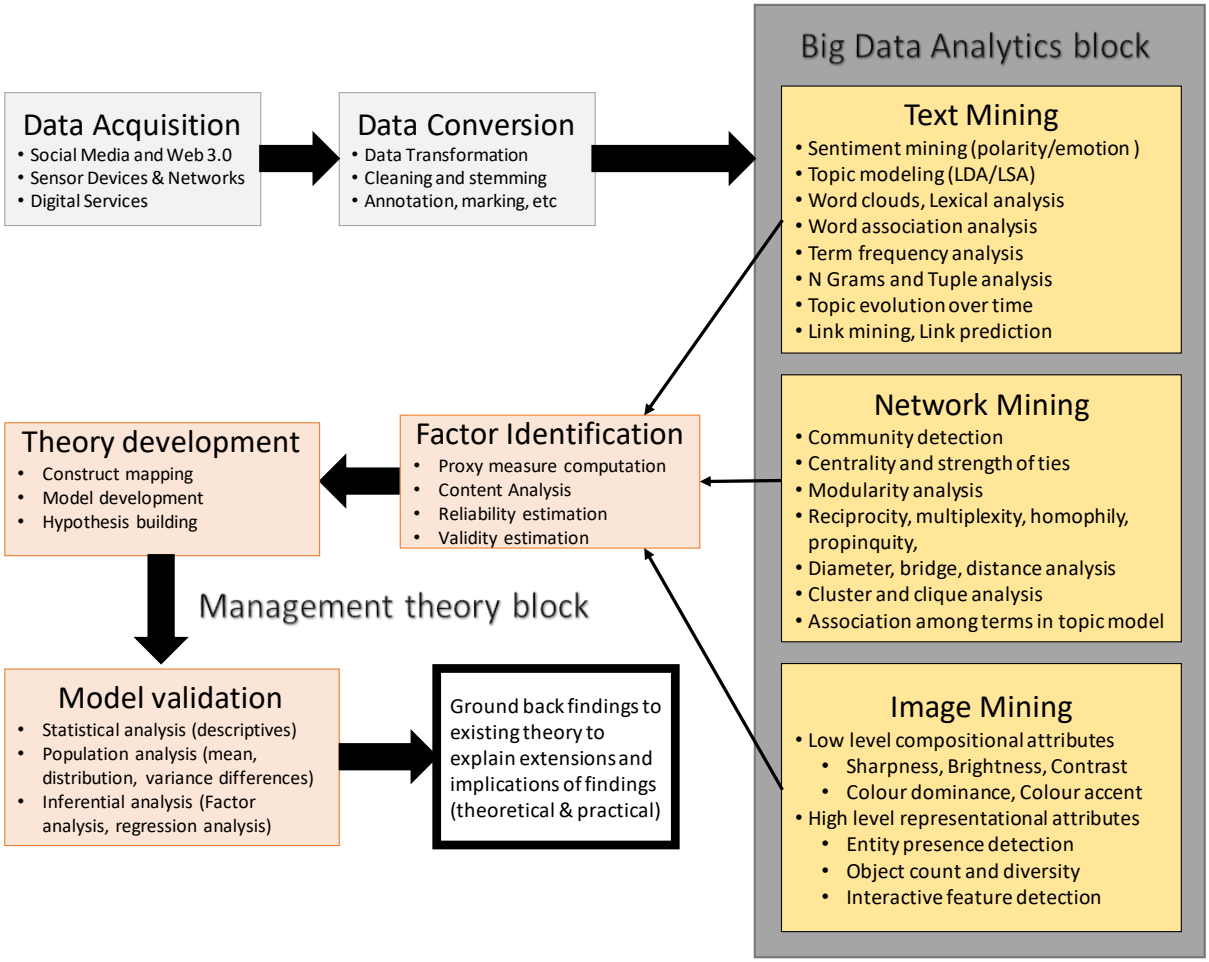

Figure 2: Process diagram with indicative detailing for theory building in big data- driven research

In the "Big Data Analytics block" in figure 2, some indicative elements have been presented based on select literature. However, there is a chance to go much deeper into this block and understand different other algorithms and approaches which may be used for theory building research. To explore that possibility, an overview of possible domains and algorithms which are documented in existing literature is analyzed based on keyword-based search in Scopus, using an approach similar to systematic literature review, as a process. The initial keyword consisted of big data and then with that when we searched Scopus, we identified other associated and relevant keywords like sentiment analysis, text mining, twitter analytics, machine learning, natural language processing, artificial intelligence, cognitive

computing, network science, network mining, image analytics, image mining and analytics as dominant author based keywords. Using these keywords, we searched for all journal publications in Scopus, in the business management and accounting and decision science for a period of five years. A total of 63,044 articles were found published between 2016 to 2020 in the categories like business management and accounting, decision science and social science across Scopus indexed journals, which capture different quality of publications. The keywords were analysed using network science to identify association rules among the keywords and major emerging themes. The results are illustrated in fig. 3.

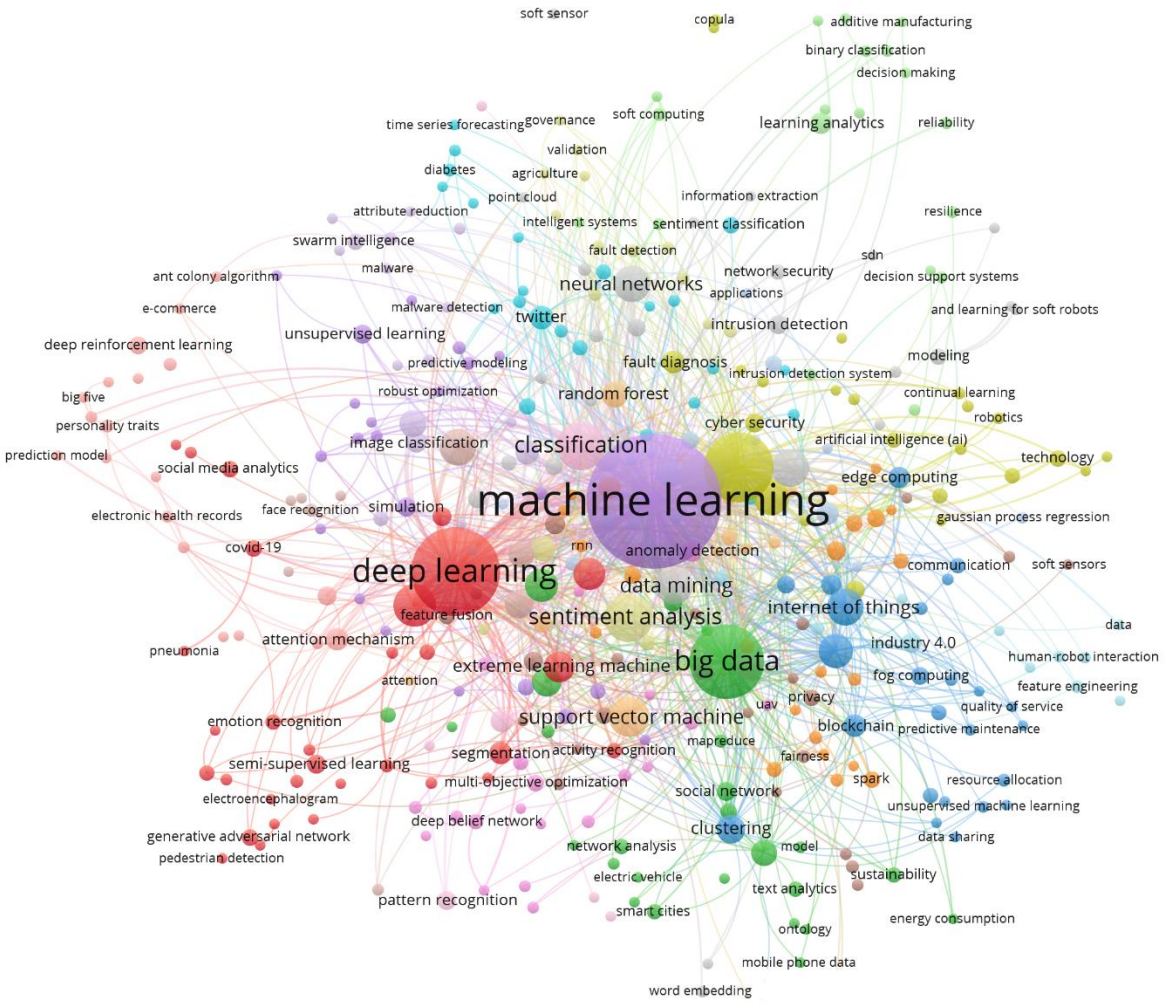

Fig 3: Keyword association among articles using data driven approaches in Scopus

The findings may be informative because it presents an overview of many algorithmic approaches which studies have already used in literature across different journals. Further, the context or the domain within which the algorithm has been used in literature is also indicated. This helps to develop ideas surrounding possible approaches, which can be adopted for specific types of data and for addressing a problem statement in a domain where it has not been explored before. The connections which are explored provide insights of established literature, although different applications of these algorithms and methods may be adopted in data driven research. More interesting may be the methods, which have not been connected to theory building in a specific domain. For example, studies on learning analytics have not been explored using neural networks, and so an application based study where consumer interaction is mined on a learning management system could be extremely informative. Another area

that has been explored relatively lesser is the domain of NeuroIS and Deep Learning applications (e.g. see Dimoka et al., 2012). Further studies on energy consumption and sustainability have not been explored extensively using sentiment analysis or text mining.

These gaps could give possible directions for future research to explore and contribute to information management theory development. Further an exploration of using a different algorithmic method on a context where a different method has been used, may also provide the scope of theory development as the data would generate different output, and so the nature of the variables used for theorizing could differ significantly. This would be relevant if some of the newer algorithmic methods were used like *deep learning*, *extreme learning* and *recurrent neural networks* to explore a study which has used data from a context to address the relationships among other computed attributes.

4. Establishing trust on the outcome

While data-driven research is gaining prominence, many of these studies rarely introspect why a phenomenon is better explained by a theory and limit the analysis to what is happening by mining data. Statistical validation of even the "what" is important as visualized information often may not be statistically significant. Many studies try to capture data and showcase applications of data science and visualization of unstructured, large volumes of data by demonstrating sentiment analysis, text mining, networks, and communities. But there is often no significant contribution to the focal theoretical context. It is important for authors to explain the context why a phenomenon is happening rather that what is happening. Units of analysis could be individuals, groups or organizations.

Possible theoretical contributions could explain the nature of interaction among entities from more than one unit-type. Multi-methodological research-based studies would be extremely insightful under such contexts where data driven approaches like social media analytics and network science may be combined with case studies, limited surveys, qualitative approaches, netnography and other approaches to provide complementary insights. Inferential analysis would be expected while developing research questions and hypothesis. Such methodological improvements are expected to contribute towards the grand wishes of management research like improving theoretical coverage, reduce interference of noise, improving validity and reliability of models, strengthening causality linkages, reduce methodological biases, and develop theory that works well within limitations (Aguinis and Edwards, 2014). Such mixed research methodologies could be extremely useful for theoretical insights and development of counter-intuitive models which are grounded in data.

Statistical validation of all proposed theoretical models is essential in all big data driven management research (George et al., 2016). This is only feasible if appropriate hypothesis development is followed during proposed model development. Many studies have demonstrated how statistical validation can be achieved. For example, approaches like topic modelling can be ratings or sentiments in an econometric model for validation (Büschken and Allenby, 2018). Further multivariate analysis has been demonstrated with logistic regression when the dependent variable is of two classes (Saxton et al., 2019). Similarly, Li and Xie (2019) demonstrate validation using a bivariate zero-inflated negative binomial econometric model for validating attributes derived both from text and images in social media posts, based on natural language processing and image processing. Similarly, Adamopoulos et al. (2018) introduces an econometric model for capturing elements like word of mouth effectiveness by modelling attributes derived out of unstructured data. It is important to specify the exact econometric model used with scientific notations for replicability of the study so that the theoretical model is testable and this falsifiable. Adding control variables and robustness tests can make the model validation more challenging but the outcome more reliable (Goh et al., 2013). Further multi-modal data analysis can

create theoretical contributions which are otherwise infeasible to validate (Balducci and Marinova, 2018).

Reporting of the findings should be aligned with the research questions and move beyond propositions which appear supported from data visualization. This will enable better objectivity in findings and contribute to theory building. A good theoretical model should attempt to touch upon all the qualities of a good theory, namely novelty, extendibility, parsimony, generalizability, reproducibility, falsifiability, internal consistency, and stand the test of time. While a study may not be able to capture all the elements of this wish list mainly due to methodological limitations of existing body of knowledge, attempts must be made to capture more elements of this list to ensure the study has greater impact.

5. Discussion

It is imperative to develop research questions which can be validated to an extent through testable hypothesis wherever applicable. This facilitates theory building and contributing to the existing body of literature. Most of the current approaches attempt to build weak propositions given the challenges of high volumes, velocity and veracity within the data. However, valuable data-driven research in IS which is generalizable and stand the test of time, would only emerge if attempts are made to develop inferential models which could be statistically validated. Our editorial is targeted to bridge this gap in existing methodologies and attempts, and provides a direction for future research and theory building. We hope that when researchers adopt the stages highlighted and follow the path specified, theoretical contributions in IS would be achieved more objectively. The rigour of such methodologies would also reduce biases, enhance falsifiability, and ensure the theoretical models stand the test of time.

5.1 Implications

The directions provided in this editorial note has many implications for the researchers working in this domain. The article highlights the stages in which researchers need to think about while designing a data driven approach. The directions attempt to bring objectivity in the research process while explaining what can be undertaken in each stage of this big data-driven research so that there is more trust and objectivity in the outcome. We highlight the need to establish reliability and validity within the research protocols itself. Further we attempt to highlight how such reliability and validity may be established by researchers who have different training background (computational science versus qualitative research competency). Further we highlight the need to move beyond descriptives and inferences from visualization, towards inferential statistics whereby theory building can be more objectively established. There is a need to follow all these steps in sequence to ensure that the final outcome would have higher theoretical contribution in the discipline of IS. The data used in these future studies may use sources like online forums, social media, devices connected through the internet of things (IoT), smartphones, wearables, digital services, location based services, and even from more than one source and data variety in multi-modal studies. Such studies would need to go a step further that being pure computational studies and contribute to the theory development, possible from mixed research methodologies, whenever needed.

For example, interaction and theory building between individuals for political contexts has been documented in IS literature for elections (Wattal et al., 2010; Grover et al., 2017; Grover et al., 2019a; Singh et al, 2020). Again, approaches like sentiment mining (Salehan and Kim, 2016) can be used to provide theoretical contribution within industries like ecommerce and, thus, contribute to management theory building. Similarly, contributions in understanding individual level impacts and usage

experience of an emerging technology based on mining UGC has been analysed and documented in literature (Grover and Kar, 2020; Kar, 2020) Such impact of interaction may be analysed between individuals and influencers who represent organizations (Grover et al., 2019b). Similarly, the contexts like adoption or diffusion of industrial technologies can be studied in organizational context based on data in social media along with other sources of data coming in from academic or trade literature (Grover et al., 2019c). Further, user generated content may be mined to understand a concept, like future of work (Sarin et al., 2020). Again, Ross et al. (2019) studied network elements to investigate the threat levels of social bots. Similarly, organizational level findings, both positive or negative consequences, could also be extremely informative in such studies, even if hypothesis testing is not done, but a lens of theoretical nature is adopted to look at a complex phenomenon (e.g. Aswani et al., 2018 used transaction cost economics). All these studies attempted to contribute to theory building in the context by mining big data which has a lot of veracity, volume and variety. The units of analysis could touch upon more than one entity: like individuals, organizations, society or policy. The data could be derived from more than one data type, namely text, images and networks. While text and images have somewhat been explored, and separately text and networks have been somewhat studied, modelling outcome of images and networks is not yet explored adequately in the current literature. The contribution would help us examine and understand the nature of interaction between these entities or stakeholders based on data-driven methodologies.

5.2 Research agenda

The area of big data-driven research is showcasing phenomenal traction among researchers which is transforming the way IS discipline had ever witnessed. The space of action research using big data is witnessing a major transformation. However these initiatives need to contribute back to theory building in information systems. With the background of this context, given the evolution of the new platforms, emerging technologies and access to data, we present some possible opportunities to future researchers to explore:

1. How can we explain user interaction, consumer experiences and impacts for emerging business models like digital services or platform economy?
2. How can big data-driven research be used to explain digital service or technology adoption, usage, and impact behaviour based on mining user generated content?
3. How can user generated content be mined to explain user behavior in socio-political contexts like opinion polarization, acculturation or communal changes?
4. How can user engagement or disengagement in digital platforms or technologies like wearables be measured and explained based on big data analytics?
5. How can we explain phenomenon surrounding digital service usage, user migration and experiences based on network data (say from telecommunication services)
6. How can we develop typology of users or organizations based on user generated content in forums, social media and platforms?
7. How can we explain relationships between organizations and other stakeholders (organizations, individuals, customers, suppliers, government, etc) based on online content in platforms and e-markets and their impacts on engagement or disengagement?
8. How can we model adverse impacts of disruptive technologies like artificial intelligence, blockchain, internet of things based on usage behaviour or user generated content?
9. How can we explain user behavior and impacts based on data derived out of sensor based data like wearables or other smart technologies used at home?

10. How can we explain community driven behaviour for information and misinformation propagation, cascade and changes to the ecosystem?
11. How can we model determinants of information quality, misinformation or disinformation based on computed text, network and user attributes?
12. How can we model computationally derived attributed of images and videos to study consumer engagement and interaction processes and outcome?
13. How can theories be developed to explain grand socio-political problems and challenges of like pandemic management, sustainable development goals, political harmony, etc?
14. How can methods of NeuroIS and facial recognition be used to explain user behaviour, traits and socio-political behavioural inclination?
15. How can multi-modal data analysis be used to create knowledge surrounding the process and impacts of use of emerging smart technologies?

Needless to say, such exploration would need to ensure that theoretical contributions are well developed into the IS discipline. These studies should explicitly attempt to address domain, socio-political, structural or ontological, and epistemological questions through the development of management theories, organization theories, behavioural theories, and systems theories.

Further it is also required to be recognised how big data-driven research would also be part of design science paradigm of IS research, whereby innovative solutions may also be created which help to define ideas, capabilities, practices and innovative products or services through big data analysis (Hevner et al., 2004). Such a lens of design science may consider data itself to be an artefact which has relevance to a problem or context, and big data methods and model validation help in the evaluation of designs created iteratively through observational, analytical, experimental, descriptive and inferential validation methodologies having adequate rigour. Research which cannot explicitly use lens of IS in their approach can also adopt the lens of design science in developing the study in such big data-driven action research.

6. Conclusion

IS research with big data is still at a nascent stage. There is a lot of scope for it to mature in the years to come, and to develop valuable theoretical contributions. We feel that the authors of research papers should attempt to address the ten points which are highlighted in this editorial note. All researchers should strive to meet the following objectives with possible aligned methodological solutions as indicated in table 1:

SN	Focused objective	Possible methodological solutions
1	Data acquisition based on "theoretical research questions" to minimize data acquisition bias.	Sampling, keyword, entity and user profile identification. Address data imbalance problems if needed.
2	Handle outliers in data better	Data cleaning, stemming, sub-sampling
3	Improve validity of measures	Qualitative intervention and inputs of subject matter experts may be required. Focus group discussions and field experiments may help.
4	Improve reliability of measures	Reporting inter-coder reliability and category reliability for content analysis type approaches.
5	Use computationally derived measures from data where ever possible in	More than one measure is a better proxy for constructs identified from literature. Hypothesis building is very important, wherever feasible.

	inferential model and bring objectivity, to the dependent variable	
6	Understand data limitations from a single type of data	Use of text, networks, images and links or a mix of these data types, for building the models would be desirable. Multi-modal data analysis would be particularly exciting and enriching.
7	Address data measurement challenges due to biases affecting the generation of the data	Using objective or computed variables which can be used as control variables, would improve trust on the outcome.
8	Minimize trade-off between internal and external validity of research model	Statistical validation of differences between groups, inferential statistics like penalised regression, logit models or multivariate analysis.
9	Check the data compatibility in measures	Time period match of data, adjusting for multi-source data problems
10	Realistic assessment of limitations and trade-offs should be reported.	Report low explainability of inferential model, if needed. Data is expected to have high noise.

Table 1: Bringing theoretical contributions in big data research methodologically

Approaching and adopting the highlighted research methodologies would bring possibilities to contribute in IS theory development. With methodological improvements, studies would also be able to minimize the usual trade-offs between internal (i.e., confidence in inferences about contextual findings) and external validity (i.e., confidence in the generalizability of findings). This would bring in more objectivity and rigour of the findings and enable big data-driven research to take the next steps beyond the "what has happened" to "why it happens". Further it is important to note that the perspective taken in this editorial review, can be extended for thinking from the perspective of design science and action research value for IS, and how these can complement data-driven studies. Future studies need to attempt to integrate these islets of literature to make theory building for more practice relevant.

References

Adamopoulos, P., Ghose, A., & Todri, V. (2018). The impact of user personality traits on word of mouth: text-mining social media platforms. *Information Systems Research*, 29(3), 612-640.

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, 51(1), 143-174.

Ali, H., & Kar, A. K. (2018). Discriminant analysis using ant colony optimization–an intra-algorithm exploration. *Procedia computer science*, 132, 880-889.

Angelopoulos, S. and Merali, Y. (2017). "Sometimes a cigar is not just a cigar: Unfolding the transcendence of boundaries across the digital and physical", *ICIS International Conference in Information Systems*, Seoul, Korea.

Arora, A., Bansal, S., Kandpal, C., Aswani, R., & Dwivedi, Y. (2019). Measuring social media influencer index-insights from Ffacebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, 49, 86-101.

Aswani, R., Kar, A. K., Ilavarasan, P. V., & Dwivedi, Y. K. (2018). Search engine marketing is not all gold: Insights from Twitter and SEOClerks. *International Journal of Information Management*, 38(1), 107-116.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, 46(4), 557-590.

Barabási, A. L. (2016). *Network science*. Cambridge university press.

Barki, H., Rivard, S., & Talbot, J. (1993). A keyword classification scheme for IS research literature: an update. *MIS Quarterly*, 17(2), 209-226.

Bharathi, S. V. (2017). Prioritizing and ranking the big data information security risk spectrum. *Global Journal of Flexible Systems Management*, 18(3), 183-201.

Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6), 953-975.

Business Frontiers. (2020). Business Strategy As The Driver: Perspective – technology transformation. https://business-frontiers.org/2019/09/08/business-strategy-as-the-driver-perspective-technology-transformation/ Accessed 15th July, 2020.

Chae, B. K. (2015). Insights from hashtag #supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, 165, 247-259.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, 20(1), 3-31.

Chakraborty, A., & Gupta, B. (2016). Paradigm phase shift: RF MEMS phase shifters: An overview. *IEEE Microwave Magazine*, 18(1), 22-41.

Chakraborty, A., & Joe, M. (2017). 7 ways how big data is changing marketing. Link: https://tech-talk.org/2017/11/28/7-ways-how-big-data-is-changing-marketing/ Accessed: 8th July, 2020.

Chakraborty A., Kar A.K. (2016) A Review of Bio-Inspired Computing Methods and Potential Applications. In: Lobiyal D., Mohapatra D., Nagar A., Sahoo M. (eds) Proceedings of the International Conference on Signal, Networks, Computing, and Systems. *Lecture Notes in Electrical Engineering*, 396, 155-161. Springer, New Delhi.

Chakraborty, A., & Kar, A. K. (2017). Swarm intelligence: A review of algorithms. In: Patnaik S., Yang XS., Nakamatsu K. (eds) Nature-Inspired Computing and Optimization. *Modeling and Optimization in Science and Technologies*, 10, 475-494. Springer, Cham.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165-1188.

Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X., & Qu, H. (2010, March). Context preserving dynamic word cloud visualization. In 2010 IEEE *Pacific Visualization Symposium* (PacificVis) (pp. 121-128). IEEE.

Curtin, J., R. J. Kauffman and F. J. Riggins (2007). "Making the most out of RFID technology: a research agenda for the study of the adoption, usage and impact of RFID." *Information Technology and Management*, 8(2): 87-110.

Dimoka, A., Davis, F. D., Gupta, A., Pavlou, P. A., Banker, R. D., Dennis, A. R., Ischebeck, A., Müller-Putz, G., Benbasat, I., Gefen, D., Kenning, P.K., Riedl, R., Brocke, J.V. & Weber, B. (2012). On the use of neurophysiological tools in IS research: Developing a research agenda for NeuroIS. *MIS Quarterly*, 36(3), 679-702.

Dong, W., Liao, S., & Zhang, Z. (2018). Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems*, 35(2), 461-487.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management*, 48, 63-71.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management*, 48, 63-71.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Kar, A.K. & Galanos, V. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dwivedi, Y. K., Rana, N. P., Janssen, M., Lal, B., Williams, M. D., & Clement, M. (2017). An empirical validation of a unified model of electronic government adoption (UMEGA). *Government Information Quarterly*, 34(2), 211-230.

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2019). Re-examining the unified theory of acceptance and use of technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers*, 21(3), 719-734.

Dwivedi, Y. K., Wade, M. R., & Schneberger, S. L. (Eds.). (2011). *Information systems theory: Explaining and predicting our digital society* (Vol. 1). Springer Science & Business Media.

Fosso Wamba, S., E.W.T. Ngai, F.J. Riggins and S. Akter (2017). "Big data and business analytics adoption and use: a step toward transforming operations and production management?" *International Journal of Operations & Production Management*, 37(1): 2-9.

Fronzetti Colladon, A., Gloor, P., & Iezzi, D. F. (2020). Editorial introduction: The power of words and networks. *International Journal of Information Management*, 51, 102031. https://doi.org/10.1016/j.ijinfomgt.2019.10.016

Gefen, D., Straub, D., & Boudreau, M. C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems*, 4(1), 1-76.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. *Academy of Management Journal*, 59(5), 1493–1507.

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. https://doi.org/10.1016/j.ijinfomgt.2019.102048

Ghani, N. A., Hamid, S., Hashem, I. A. T., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior*, 101, 417-428.

Gloor, P., Fronzetti Colladon, A., Giacomelli, G., Saran, T., & Grippa, F. (2017a). The impact of virtual mirroring on customer satisfaction. *Journal of Business Research*, 75, 67–76.

Gloor, P., Fronzetti Colladon, A., Grippa, F., & Giacomelli, G. (2017b). Forecasting managerial turnover through e-mail based social network analysis. *Computers in Human Behavior*, 71, 343–352.

Goh, K. Y., Heng, C. S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research*, 24(1), 88-107.

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

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(3), 203-229.

Grover P., Kar A.K., Dwivedi Y.K., Janssen M. (2017) The Untold Story of USA Presidential Elections in 2016 - Insights from Twitter Analytics. In: Kar A. et al. (eds) Digital Nations – Smart Cities, Innovation, and Sustainability. I3E 2017. *Lecture Notes in Computer Science*, 10595, 339-350. Springer, Cham.

Grover, P., & Kar, A. K. (2020). User engagement for mobile payment service providers–introducing the social media engagement model. *Journal of Retailing and Consumer Services*, 53, 101718. https://doi.org/10.1016/j.jretconser.2018.12.002

Grover, P., Kar, A. K., & Ilavarasan, P. V. (2019b). 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, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019a). 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., Janssen, M., & Ilavarasan, P. V. (2019c). Perceived usefulness, ease of use and user acceptance of blockchain technology for digital transactions–insights from user-generated content on Twitter. *Enterprise Information Systems*, 13(6), 771-800.

Grover, P., Kar, A.K. & Dwivedi, Y.K. (2020). Understanding Artificial Intelligence Adoption in Operations Management – Insights from the review of academic literature and social media discussions. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03683-9

Grover, V. (2020). Do We Need to Understand the World to Know It? Knowledge in a Big Data World. *Journal of Global Information Technology Management*, 23(1), 1-4.

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), 268-291.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Ting, L., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77, 354-377.

Hannigan, T. R., Haans, R. F., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M.S., Tchalian, H., Wang, M.S., Kaplan, S. and P. Devereaux Jennings Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. *Academy of Management Annals*, 13(2), 586-632.

He, W., Zhang, W., Tian, X., Tao, R., & Akula, V. (2019). Identifying customer knowledge on social media through data analytics. *Journal of Enterprise Information Management*. 32(1), 152-169.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124, 226-251.

Hevner, A. R., et al. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75-105.

Joseph, N., Kar, A. K., Ilavarasan, P.V., & Ganesh, S. (2017). Review of discussions on internet of things (IoT): Insights from twitter analytics. *Journal of Global Information Management*, 25(2), 38-51.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7(2), 1-34.

Kar, A. K. (2015). Integrating websites with social media–An approach for group decision support. *Journal of Decision Systems*, 24(3), 339-353.

Kar, A. K. (2016). Bio inspired computing–a review of algorithms and scope of applications. *Expert Systems with Applications*, 59, 20-32.

Kar, A. K. (2017). Apache Spark – the next big thing for Big Data. Link: https://www.business-fundas.com/2017/apache-spark-the-next-big-thing-for-big-data/ Accessed: 15th July 2020.

Kar, A.K. (2020). What affects Usage Satisfaction in Mobile Payments? Modelling User Generated Content to develop the "Digital Service Usage Satisfaction Model". *Information Systems Frontiers*. DOI: 10.1007/s10796-020-10045-0

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: a systematic literature review through text mining. *IEEE Access*, 8, 67698-67717.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research*, 4(1), 8-18.

Kaushik, K., Mishra, R., Rana, N. P., & Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on Amazon. *Journal of Retailing and Consumer Services*, 45, 21-32.

Kumar, A., Singh, J. P., Dwivedi, Y. K., & Rana, N. P. (2020). A deep multi-modal neural network for informative Twitter content classification during emergencies. *Annals of Operations Research*, 1-32.

Lambrecht, A., & Tucker, C. (2013). When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research*, 50(5), 561-576.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Lee, I. (2018). Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons*, 61(2), 199-210.

Lee, U., Han, K., Cho, H., Chung, K. M., Hong, H., Lee, S. J., Noh, Y., Park, S., & Carroll, J. M. (2019). Intelligent positive computing with mobile, wearable, and IoT devices: Literature review and research directions. *Ad Hoc Networks*, 83, 8-24.

Lyytinen, K. (2009). Data matters in IS theory building. *Journal of the Association for Information Systems*, 10(10), 715-720.

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, 19(12), article 1. DOI: 10.17705/1jais.00526

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2), 293-334.

Mishra, N., & Singh, A. (2018). Use of twitter data for waste minimisation in beef supply chain. *Annals of Operations Research*, 270(1-2), 337-359.

Misirlis, N., & Vlachopoulou, M. (2018). Social media metrics and analytics in marketing–S3M: A mapping literature review. *International Journal of Information Management*, 38(1), 270-276.

Oh, O., Eom, C., & Rao, H. R. (2015). Role of social media in social change: An analysis of collective sense making during the 2011 Egypt revolution. *Information Systems Research*, 26(1), 210-223.

Paice, C. D. (1994). An evaluation method for stemming algorithms. In SIGIR'94 (pp. 42-50). Springer, London.

Pentland, A. (2008). Honest Signals: How They Shape Our World. MIT Press. USA.

Rathore, A. K., Kar, A. K., & Ilavarasan, P. V. (2017). Social media analytics: Literature review and directions for future research. *Decision Analysis*, 14(4), 229-249.

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4), 394-412.

Roy, P. K., Ahmad, Z., Singh, J. P., Alryalat, M. A. A., Rana, N. P., & Dwivedi, Y. K. (2018). Finding and ranking high-quality answers in community question answering sites. *Global Journal of Flexible Systems Management*, 19(1), 53-68.

Salehan, M., & Kim, D. J. (2016). Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems*, 81, 30-40.

Sarin, P., Kar, A. K., Kewat, K., & Ilavarasan, P. V. (2020). Factors affecting future of work: Insights from Social Media Analytics. *Procedia Computer Science*, 167, 1880-1888.

Saumya, S., Singh, J. P., & Dwivedi, Y. K. (2019). Predicting the helpfulness score of online reviews using convolutional neural network. *Soft Computing*, 1-17.

Saumya, S., Singh, J. P., Baabdullah, A. M., Rana, N. P., & Dwivedi, Y. K. (2018). Ranking online consumer reviews. *Electronic Commerce Research and Applications*, 29, 78-89.

Saxton, G. D., Gómez, L., Ngoh, Z., Lin, Y. P., & Dietrich, S. (2019). Do CSR messages resonate? Examining public reactions to firms' CSR efforts on social media. *Journal of Business Ethics*, 155(2), 359-377.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Shiau, W. L., Dwivedi, Y. K., & Lai, H. H. (2018). Examining the core knowledge on Facebook. *International Journal of Information Management*, 43, 52-63.

Shiau, W. L., Dwivedi, Y. K., & Yang, H. S. (2017). Co-citation and cluster analyses of extant literature on social networks. *International Journal of Information Management*, 37(5), 390-399.

Singhal, H., & Kar, A. K. (2015). Information security concerns in digital services: Literature review and a multi-stakeholder approach. In 2015 *International Conference on Advances in Computing, Communications and Informatics*, pp. 901-906. IEEE.

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. K. (2019). Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, 283, 737–757

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, 70, 346-355.

Singh, P., Dwivedi, Y. K., Kahlon, K. S., Pathania, A., & Sawhney, R. S. (2020). Can twitter analytics predict election outcome? An insight from 2017 Punjab assembly elections. *Government Information Quarterly*, 37(2), 101444.

Subramanian, A., & Nilakanta, S. (1994). Measurement: a blueprint for theory-building in MIS. *Information & Management*, 26(1), 13-20.

Sutton, R. I., & Staw, B. M. (1995). What theory is not. *Administrative Science Quarterly*, 371-384.

Tirunillai, S., & Tellis, G. J. 2012. Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, 31(2), 198–215

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28, 3–27.

Wagner, K. (2020). Facebook Expands Location Data Sharing With Covid-19 Researchers. Bloomberg. Link: https://www.bloomberg.com/news/articles/2020-04-06/facebook-expands-location-data-sharing-with-covid-19-researchers Last accessed: 13th June, 2020.

Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd *International Conference on Machine learning* (pp. 977-984).

Wattal, S., Schuff, D., Mandviwalla, M., & Williams, C. B. (2010). Web 2.0 and politics: the 2008 US presidential election and an e-politics research agenda. *MIS Quarterly*, 669-688.

Weick, K. E. (1989). Theory construction as disciplined imagination. *Academy of Management Review*, 14(4), 516-531.

Weick, K. E. (1995). What theory is not, theorizing is. *Administrative Science Quarterly*, 40(3), 385-390.

Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of management review*, 14(4), 490-495.

Wu, L. (2013). Social network effects on productivity and job security: Evidence from the adoption of a social networking tool. *Information Systems Research*, 24(1), 30-51.

Wu, P., Li, X., Shen, S., & He, D. (2020). Social media opinion summarization using emotion cognition and convolutional neural networks. *International Journal of Information Management*, 51, 101978. https://doi.org/10.1016/j.ijinfomgt.2019.07.004

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, 58, 51-65.

Zuo, Meihua, Z., and Angelopoulos, Spyros, A., and Ou, C.X., Carol Xiaojuan , C., and Liu, Hongwei, L., and& Liang, Zhouyang, L. (2020). Identifying Dynamic Competition in Online Marketplaces Through Consumers' Clickstream Data (April 22, 2020). Available at SSRN: http://dx.doi.org/10.2139/ssrn.3598889