A Review of Big Data Predictive Analytics in Information Systems Research

Alhassan Ohiomah
aohio100@uOttawa.ca

Pavel Andreev
andreev@Telfer.uOttawa.ca

Morad Benyoucef
benyoucef@Telfer.uOttawa.ca

Telfer School of Management
University of Ottawa
Ottawa, ON K1N 6N5, Canada

Abstract

Big data, with its inherent complexity, introduces new challenges for traditional business intelligence and analytics tools, and offers opportunities for organizations to use advanced solutions to exploit their highly complex data. Moreover, the use of predictive analytics on big data has emerged as an important topic for researchers and practitioners from various disciplines. This study conducts a review of the Information Systems (IS) literature on big data predictive analytics to identify the areas of big data predictive analytics that have been studied and are still in need of more research focus, and proposes specific research questions for future investigation. Overall, we found that the emergence of big data has changed the role of predictive analytics from activities such as theory generation and validation to more data-driven discovery of complex patterns and relationships between variables, and assessing the likelihood of occurrence of relationships between a dataset's variables. The outcomes of this research contribute to the IS literature by helping identify research gaps, approaches, and emerging directions in big data predictive analytics, and enable practitioners to understand the potentials and applications of this new and important concept.

Keywords: Big data, Predictive analytics, Business Intelligence, Systematic Review

1. INTRODUCTION

Organizations are witnessing a rapid growth in the volume of data they generate daily (Watson, 2014). Recent reports indicate that 4 Zettabytes (4 Trillion Gigabytes) of digital data are created every day (Goes, 2015). IBM reports that 90% of the data in the present day have been generated in the last two to three years (IBM, 2015). What poles apart is that these high volumes of data are of different variety, have different veracity, originate with different velocity and offer different values, a concept now generally known as "Big data" (Goes, 2015; Power, 2013; Shim, French, Guo, & Jablonski, 2015). Usually, organizations turn to their data to explore the challenges and opportunities existing within their business. However, although the emergence of big data offers organizations ample opportunities (Phillips-Wren, Iyer, Kulkarni, & Ariyachandra, 2015), many organizations still lack an understanding of how to better utilize these growing amounts of data to their advantage (Bedeley, 2014; Koronios, Gao, & Selle, 2014; Power, 2013). This

is because business intelligence and analytics tools used by organizations are not usually sufficient to handle the complexity of big data (Chen, Chiang, & Storey, 2012; Watson, 2014). Big data requires the application of advanced analytical techniques (Chen et al., 2012; Phillips-Wren et al., 2015; Watson, 2014; Wixom et al., 2014). In view of that, organizations are being compelled to exploit the potential of predictive analytics as well as other advanced business intelligence and analytics tools to help them unravel insights from their big data (Chen et al., 2012; Deka, 2014; Gualtieri, Rowan Curran, TaKeaways, & To, 2013; Kiron & Shockley, 2011; LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011; Watson, 2014).

Predictive analytics include methods that scan data for correlations, trends, and patterns to discover insights and make predictions of possible outcomes (Abbott, 2014; Delen & Demirkan, 2013; Kotu & Deshpande, 2014; Watson, 2014). With predictive analytics, informed decisions are made through a blend of data, analysis, and scientific reasoning rather than just human instincts or beliefs (Nettleton, 2014). Unarguably, predictive analytics have been available for a while, predominantly as a method for validating empirical models with small datasets gathered mostly from surveys and interviews (Shmueli & Koppius, 2011). Nevertheless, the emergence of big data has increased the promise of predictive analytics mainly because the latter are more effectual on multifarious large amounts of data (Moeyersoms & Martens, 2015). Predictive analytics are rapidly growing because of the shift from prevailing Business Intelligence tools to advanced analytics techniques and the massive surge of structured and unstructured data (Finlay, 2014; Kotu & Deshpande, 2014).

The use of predictive analytics on big data has emerged as an important area of study for both researchers and practitioners across various disciplines, including biosciences, medicine, computer science and engineering (Deka, 2014; Sun, Zou, & Strang, 2015). While several scholars have addressed specific research questions or built predictive models for specific applications, no far-reaching research agenda has been developed to understand what has been accomplished, how it has been accomplished and what remains to be accomplished when using predictive analytics on big data, particularly in the Information Systems (IS) literature. Shmueli and Koppius (2011) provide an understanding of the role of predictive analytics and the need to integrate it in IS research but the concept was not investigated from a big data perspective. Their review found that predictive analytics was mostly applied to small data usually gathered through surveys. Additionally, papers reviewed by Shmueli and Koppius (2011) were published between 1990 and 2006. To the best of our knowledge, no study has since then investigated the use of predictive analytics in IS research, particularly in the realm of big data.

In light of all this, we believe that the literature can benefit from an investigation of the current state and role of big data predictive analytics (BDPA) in IS research. There is now a sizable body of research to be reviewed starting from 2006. Hence, there is a need to synthesize the literature to determine what has been done and what is missing in this area. Also, it will be interesting to uncover whether the era of big data has changed the type and context of predictive analytics research conducted within the IS field, and whether the complexity of big data has introduced new predictive models, algorithms and application domains. This paper aims to do just that. The current study makes three main contributions.

- First, we review the interplay between big data, analytics and business intelligence and provide a definition of the term "Big data predictive analytics (BDPA)" (Background section). We believe that this is the first work that combines unique concepts from the literature to offer a cohesive definition of the term.
- Second, we conduct a structured review of the academic literature on BDPA (Research method section) and reveal insights on research contexts, topics and applications of BDPA (Analysis section). For instance, we found that the majority of the reviewed studies used techniques that were not frequently employed for predictive modelling before the era of big data.
- Third, our discussion will help researchers understand the current body of knowledge, identify key gaps in the literature on BDPA, and suggest several questions that can serve as a starting point for further research in this area (Discussion section).

2. BACKGROUND

To provide an understanding of big data, analytics and business intelligence in the context of decision-making, this section describes the three concepts and how they relate to each other. We

also propose a definition of BDPA through a synthesis of definitions in the literature.

In the beginning, the concept of big data denoted large volumes of data. The financial industry (i.e., Stock markets, Credit institutions) has been dealing with such voluminous data since the late 1990s. Over time, information and communication technologies promoted an environment where large amounts of data were easy to collect from different sources at different speeds. The sources of such data include sensors of various kinds, social media posts, digital pictures and videos, purchase transaction records, and cell phone GPS signals (IBM, 2015). Yet, scholars suggest that many data sources today remain untapped or underutilized (Franks, 2012; Watson, 2014). The size, diversity and delivery speed of big data creates huge challenges for organizations. Such challenges involve the viability of traditional business intelligence and analytics tools, as well as the opportunities for organizations to employ cutting-edge tools to help them obtain optimum value from their highly complex data. Befittingly, research on big data, analytics and business intelligence has received growing attention from the academic community in the past few years (Chen et al., 2012; Phillips-Wren et al., 2015; Watson, 2014). Next, we discuss the relationship between big data, analytics and business intelligence.

Big Data

Watson (2014, p. 1249) defined big data as *"data that is high volume, high velocity and or high variety which requires new technologies and techniques to capture, store, and analyze it and is used to enhance decision making, provide insight and discovery and support and optimize processes".* Two other dimensions (i.e., Veracity and Value) have been used to characterize big data (Shim et al., 2015). The dimensions of big data offer opportunities for insight, but a real challenge is how to turn big data into valuable insights. Organizations constantly gathering big data do not directly create business value; value is created only when big data is analyzed and utilized for decision making (Watson, 2014).

Analytics

Analytics involve the use of iterative and methodical techniques to discover, analyze and interpret meaningful patterns from data (Baltzan & Welsh, 2015). Analytics support businesses with technologies needed to analyze data, visualize it and create models to foresee future problems and opportunities, and tools to optimize business processes (Delen & Demirkan, 2013). **Big Data Analytics (Big Data + Analytics)** is a concept used to describe the analytics of big data (Chen et al., 2012; Sun et al., 2015). Analytics build on principles from data mining, statistical analysis and operations research (Chen et al., 2012). There are currently three core categories of analytics, namely Descriptive, Predictive and Prescriptive analytics (Deka, 2014; Delen & Demirkan, 2013; Watson, 2014). In this study, we only focus on Predictive analytics. **Predictive analytics** include methods that investigate historical and current data for hidden patterns and relationships to predict future trends and outcomes (Shim et al., 2015). Predictive analytics reveal insights on "what will happen" and "why it will happen" (Deka, 2014; Delen & Demirkan, 2013). Predictive analytics by design include key aspects of descriptive and prescriptive analytics as well (Hair Jr, 2007). It uncovers relationships and patterns within data to forecast possible outcomes for decision optimization.

Business Intelligence

Watson (2009, p. 491) defined Business intelligence as a *"broad category of applications, technologies, and processes for gathering, storing, accessing and analyzing data to help business users make better decisions"*. The concept includes technology, systems, practices and applications that analyze business data to help organizations understand their business and market (Lim, Chen, & Chen, 2013). The term ***Business Intelligence and Analytics (Business Intelligence + Analytics)*** gained popularity and was widely adopted in the early 2000s because of the notion that business intelligence was heavily dependent on analytics (Lim et al., 2013). Chen et al. (2012) defined Business Intelligence and Analytics as "*techniques, technologies, systems, practices, methodologies, and applications that analyze critical business data to help an enterprise better understand its business and market and make timely business decisions"*. To simplify, business analytics provide insights from business data to support intelligence for smart decisions making. Thus, business analytics is essential to gain business intelligence. Together they provide tools to convert business data into information into knowledge for better wisdom, actions and understanding of a business. With a clear understating of the concepts and interplay between big data, analytics and business intelligence, we can now investigate the focal point of our research "Big Data Predictive Analytics".

Big Data Predictive Analytics (BDPA)

The era of big data provides an avenue to process highly accurate forecasts and therefore creates new application possibilities for predictive analytics (Gualtieri et al., 2013). Simply put, BDPA is predictive analytics for big data (Sun et al., 2015). As an emerging research area, not much effort has been dedicated to explicitly define BDPA. To fill this gap, we identify distinctive descriptions of big data and predictive analytics separately.

As discussed earlier, prior research has branded big data as data with 3 key dimensions namely volume, variety and velocity (Beyer & Laney, 2012; Chen et al., 2012; Watson, 2014). Additionally, veracity (Claverie-Berge, 2012; Lukoianova & Rubin, 2014) and value (Hashem et al., 2015; Lycett, 2013) were introduced as new dimensions. Hence, big data can be referred to as data with high volume, variety, velocity, veracity and value (Abbasi, Sarker, & Chiang, 2016; Gandomi & Haider, 2015; Shim et al., 2015). Notwithstanding the differences in perceptions about the meaning of predictive analytics in the literature, there is a close unanimity that whatever definition is adopted, it involves the idea of discovery of trends, relationships and patterns from data for decision making and prediction of future events (Deka, 2014; Goul, Balkan, & Dolk, 2015; Hair Jr, 2007; Kridel & Dolk, 2013; Russell, 2015; Shim et al., 2015; Shmueli & Koppius, 2011; Watson, 2014). Russell (2015) featured identification of risks and opportunities in describing predictive analytics. Similarly, Zeng (2015) featured "prediction of future events in a wide range of application contexts, as well as individual, group, societal behaviors and actions" in describing predictive analytics. As highlighted previously, analytics only involves the use of iterative and methodical techniques to discover, analyze and interpret meaningful patterns from the data (Baltzan & Welsh, 2015). Hence, predictive analytics can be referred to as the use of iterative and methodical techniques that collect and analyze data to reveal trends, relationships and patterns within it to identify problems and opportunities, predict future events, and guide decision making in a wide range of application contexts, including individual, group, and social behaviors and actions. Based on the above discussion, we offer the following definition:

> *Big data predictive analytics is the use of iterative and methodical techniques that collect, analyze, and interpret high volume, variety, velocity, veracity and value data to reveal trends, relationships and patterns within data to identify problems and opportunities, predict future events, and guide decision making in a wide range of application contexts, including individual, group, and social behaviors and actions.*

BDPA will have a profound impact in helping business organizations deal with high volumes of structured and unstructured data to generate insights that guide day-to-day operations, improve decision making and define future strategies (Deka, 2014). Next, we investigate the IS literature for published studies on BDPA to understand its current state, application

3. RESEARCH METHOD

To understand the present state of BDPA research and identify future research directions, we review the literature for relevant publications within the IS discipline. We adopt Levy and Ellis (2006) guidelines for conducting a systematic literature review. The guidelines suggest that a review of the literature should follow the inputs, processing and outputs phases. Accordingly, we identify BDPA studies from the top ranked IS journals (input) to comprehend the concept's development over time. Second, we analyze and classify relevant studies (processing). Third, we discuss the applications and state of current practice of BDPA based on the identified studies (output).

Review Inputs

A methodical search of the literature was conducted for published studies with any of the keywords "Predict*", "Forecast*", "Data driven", "data mining", "machine Learning", "Analysis" or "Analytic*" within their title, abstract and keywords. We also required that "Big data" or "Large data*" be mentioned somewhere in the content of the papers. We assume these keywords will be in the title, abstract and keywords of any literature relevant for our study. However, it is possible that our search might neglect other relevant studies that do not have these keywords in their title, abstract and keywords.

This review covers related studies published from January 2006 to June 2017. The search was conducted on top IS senior scholar basket journals as recognized by the Association of Information Systems' and Peffers and Ya (2003). We only focused on papers from top IS senior scholar basket journals because of their profound

impact and quality publications. A total of 341 studies were identified from the search and were saved to a reference manager (Endnote).

Journals	Selected	5 Year Impact Factor
Decision Support Systems	47	4.29
MIS Quarterly	10	12.22
IS Research	7	4.79
Journal of IT	5	6.95 (2016)
Journal of MIS	4	2.35 (2016)
IS Journal	3	2.82
Journal of Strategic Information Systems	2	4.61
European Journal of IS	1	2.81 (2016)
Journal of AIS	1	2.01 (2016)
Total	**80**	

Table 1: BDPA Studies Reviewed from Different IS Journals

Applicability of literature: We scrutinized the contents of these 341 studies against the following criteria to make sure they are applicable to our research; (1) Are the studies focused on prediction? (2) Are the studies big data oriented (i.e., the Data used in the study have Volume, Variety and or Velocity)? (3) Are the studies methodologically grounded (i.e., Analysis goal, Data collection, Modelling method, Validation method)? (4) Are the studies practically or theoretically relevant? Only studies that met all four criterial were selected, including a few because of their conceptual significance. Here, we excluded papers whose concepts of BDPA did not fall within the scope, such as adoption related papers e.g., (Agrawal, 2015; Li, Wu, Liu, & Li, 2015). Additionally, we left out predictive analytics studies that used a relatively small sample of data and or non-complex data to validate their predictive models e.g., Zheng et al. (2015). Also, discussion notes and some non-influential non-empirical papers were excluded. This resulted in a final list of 77 relevant studies. Additionally, a review of the references of these 77 papers and a forward reference search (i.e., articles that cite articles under review) through google scholar yielded a final list of 80 relevant studies. Table 1 illustrates where the selected studies where published. Decision Support Systems (DSS) published the majority (47) of the relevant BDPA studies. DSS is a good fit for BDPA studies because of its aligned goal of supporting and optimizing the decision-making process. Another reason may be because, compared to other journals in the basket, DSS has a fast publication timeframe, which explains the high number of publications on the topic.

Publication Trend

We further examine the longitudinal trends of BDPA studies. We grouped all studies published prior to 2010 together as "Before 2010. Figure 1 shows the publication trends of our search results providing an understanding of the advancement of BDPA research over the years. We can see that from 2014, BDPA started receiving attention and the number of publications with that theme skyrocketed, with approximately 89% of the studies published between 2014 and 2017. It should be noted that this search was conducted in June 2017, thus, we suggest that no assumption should be made about the downwards curve in trend from 2016 to 2017. This may suggest that the literature will be flushed with more studies on BDPA in the coming years.

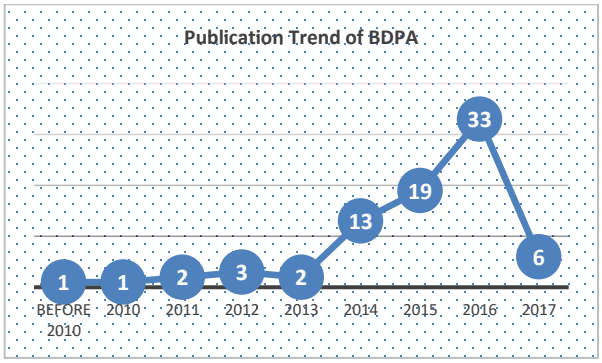

Figure 1: Publication Trend of BDPA in IS Research

4. ANALYSIS AND RESULTS

Research Category	Count	%
Empirical Research	60	75%
General Overview	10	12.5%
Privacy Issues with BDPA	4	5%
Business Value of BDPA	3	3.75%
Literature Surveys	3	3.75%
Total	80	100%

Table 2: BDPA Themes Grouped in IS research

The results of our review suggest that BDPA were mainly used in IS research for a priori data-driven discovery of relationships between variables and an assessment of the likelihood of occurrence of

the relationships between variables in the dataset. In Table 2, we see that about 75.3% of the studies are empirical in nature and the remaining 24.7% are non-empirical, exploring wide-ranging themes that require an understanding of the concept under study. These themes include a general overview (10), privacy issues of BDPA (4), business value of BDPA (3) and literature surveys (3). Table 3 (Appendix A) summarizes the selected BDPA studies.

Empirical Research

The use of BDPA in IS research is summarized in Table 4 (Appendix B) in terms of big data characteristics, data size, data source, method of analysis and application domain. We noticed that 45 of the 60 empirical studies where published in DSS. The results are further analyzed.

Big Data Characteristics*:* As illustrated earlier, big data is by design of large volume, different variety and generated at different frequencies. Our analysis reveals that 15 of the empirical studies use datasets that we identify as either high volume e.g., (Cresci, Di Pietro, Petrocchi, Spognardi, & Tesconi, 2015) or high variety (Tsai & Chen, 2014). The majority of studies (31) used data with volume and variety (Huang, Chen, & Chen, 2016; Martens & Provost, 2014), volume and velocity e.g., (Langseth & Nielsen, 2015; Moeyersoms & Martens, 2015) or velocity and variety (Dag, Topuz, Oztekin, Bulur, & Megahed, 2016; Sahoo, singh, & Mukhopadhyay, 2012). While 14 other studies use datasets that have volume, variety and velocity (Wattal, Telang, Mukhopadhyay, & Boatwright, 2011; Wu, Huang, Song, & Liu, 2016). Interestingly, we found that 9 of the studies that used datasets with the 3Vs characteristics were published between 2016 and 2017 alone. This indicates that upcoming studies on BDPA are more likely to feature datasets with the 3Vs characteristics.

Data Size: Of the empirically conducted studies that we analyzed, only 9 reported using a dataset with less than 10,000 observations. Most studies in this category analyze either text documents which mostly dwell on high dimensionality (Tsai & Chen, 2014) or health records (Wimmer, Yoon, & Sugumaran, 2016) where the number of observations is usually small because some health cases like cancer are not widely dispersed. We found that 19 of the reviewed studies used datasets ranging between 10,000 and 100,000 observations. Another 16 studies used datasets ranging between 100,000 and 1 million observations, while 7 used datasets ranging between 1 million to 10 million observations. Another 9 of the studies used datasets that contain 10 million observations or more. It should be noted that there are some studies that used multiple datasets of different size for their investigation (Langseth & Nielsen, 2015) so, we only indicated the highest number of data used in each research. Evidently, the vast amount of data available today seems to be underutilized or unavailable to the IS literature.

Data Sources: User generated content via reviews, ratings and social media has been the most exploited source of data available to BDPA in IS research with a total of 26 studies reporting their usage. Studies in this group rely on user generated content to understand user sentiments (Stieglitz & Dang-Xuan, 2013) or user preferences for recommender systems (Chen, Shih, & Lee, 2016a), with the exception of Cresci et al. (2015) who used social media data to identify fraudulent twitter followers. Another 9 studies used historical transactional data about customers in their study, such as (Carneiro, Figueira, & Costa, 2017). Another 7 studies report using health records for their investigation. Additional 6 studies used datasets other than the popular sources outlined. Datasets used in these 6 studies where collected from police theft reports (Camacho-Collados & Liberatore, 2015), lake data (Jiang, Liu, Zhang, & Yuan, 2016), or multiple sources e.g., (Bogaert, Ballings, & Van den Poel, 2016; Geva, Oestreicher-Singer, Efron, & Shimshoni, 2017; Pai, Wu, & Hsueh, 2014). Data used by 11 other studies were collected via text documents (3), email or text messages (2), census data (3) and website content (4). Our analysis indicates that IS studies are making use of more publicly available data. A reason for this might be the ethical constrains involved in collecting institutional data, data privacy considerations, or the fear that revealing data might affect competitive advantage.

Analysis Techniques: It is important to note that most of the studies we reviewed report using multiple modelling techniques for their analysis, hence we only documented the techniques that yielded the best performance. Our analysis shows that a majority (23) of the reviewed studies used techniques that were not frequently used for predictive modelling before the era of big data. For instance, Huang et al. (2016) used a Google similarity distance measure to suggest a recommender system. Another example is Khairul and Shahrul (2015) who introduced an identity matching model using Q-gram indexing. 9 studies report using regression models for their analysis e.g., (Bardhan, Oh, Zheng, & Kirksey,

2014). 5 studies report using Bayesian models based on networks (Coussement, Benoit, & Antioco, 2015; Wattal et al., 2011) or hidden Markov models (Jiang et al., 2016; Sahoo et al., 2012). Another 4 studies report using decision tree models. Interestingly, 3 of those studies where about evidence based medicine (Dag et al., 2016; Gómez-Vallejo et al., 2016; Meyer et al., 2014). This is because decision tree models are suitable for problems with sequences of what-if scenarios that can lead to various outcomes. Medical decisions are an example of such problems since health practitioners are continually faced with situations where they make crucial decisions to determine the right diagnosis, the ideal treatment or the survival chances of patients. Only 3 studies report on new algorithms that manage the complexity of the big data they had to investigate. For instance, Tsai and Chen (2014) introduced an efficient genetic algorithm for reducing high dimensional data. Also, 3 studies report using support vector machines for their investigations. Finally, 2 studies each report using matrix factorization, naïve Bayes, neural networks, rough sets and times series techniques for their data analysis. This suggests that longstanding predictive analytics techniques have been used in the literature for prediction using big data.

Application Domain: Among the IS studies analyzed, 11 were conducted to understand and predict the sentiment of users about subjects such as movies (Fersini, Messina, & Pozzi, 2014) and products (Salehan & Kim, 2016). Another 11 studies were conducted to develop recommender systems for movies, products, or predict uncertainty (Banerjee, Bhattacharyya, & Bose, 2017; Zhang, Guo, & Chen, 2016) e.t.c. Also, 10 other studies report using BDPA in fields such as predicting event attendance (Bogaert et al., 2016), forecasting microsystem in biological and disease control (Jiang et al., 2016) and generic fields (Tsai & Chen, 2014). Additional 9 studies used BDPA to gather market intelligence for segmenting e.g., (Wattal et al., 2011), sales lead qualification (D'Haen, Van den Poel, Thorleuchter, & Benoit, 2016), or better targeting consumers e.g., (De Cnudde & Martens, 2015; Moeyersoms & Martens, 2015; Pournarakis, Sotiropoulos, & Giaglis, 2017). Also, 6 other studies each where applied in health domain support medical diagnosis e.g., (Gómez-Vallejo et al., 2016) or index personal health profiles e.g., (Bardhan et al., 2014). Extra 6 studies where applied to anomaly and fraud detection in issues such as identifying fraud twitter accounts (Cresci et al., 2015) and identifying phishing for internet fraud (Abbasi et al., 2015). An additional 4 studies where applied to financials to predict firm value for stock boosting purposes (Luo & Zhang, 2013; Shynkevich, McGinnity, Coleman, & Belatreche, 2016) or to determine crowdfunding outcomes (Yuan, Lau, & Xu, 2016). 2 studies applied BDPA to identify defective toys (Winkler, Abrahams, Gruss, & Ehsani, 2016) or predict crime occurrence (Camacho-Collados & Liberatore, 2015). Only 1 study used text mining to classify similar documents (Martens & Provost, 2014). This review suggests that predictive analytics has been widely recognized and utilized by several industries to unravel insights from their big data.

General Overview

Overview studies are key to understanding the capabilities and issues associated with contemporary research areas. Earlier research focused on the general description and introduction of BDPA. Three research topics were investigated in this area namely big data issues and challenges (Clarke, 2016; Constantiou & Kallinikos, 2014), research directions of big data analytics (Abbasi et al., 2016; Phillips-Wren et al., 2015), theory and societal implications of big data analytics (Chang, Kauffman, & Kwon, 2014; Newell & Marabelli, 2015). Six additional editorials, commentary, issues and opinions publications on topics related to big data predictive analytic where also added to these category (Agarwal & Dhar, 2014; Bhimani, 2015; Markus, 2015; Müller, Junglas, Brocke, & Debortoli, 2016; Sharma, Mithas, & Kankanhalli, 2014; Woerner & Wixom, 2015; Yoo, 2015).

Privacy Related Research on BDPA

There is a growing concern by organizations and end users about the privacy implications of big data analytics. However, not much research has been conducted on how best to manage and prevent sensitive information disclosure. In fact, the same authors investigated 3 of the 4 studies reviewed on this topic. First, Li and Sarkar (2006) offered a perturbation method that organizations can use to prevent or limit the disclosure of sensitive information on categorical data when used during classification analysis. To address the issue of identity matching that can lead to privacy violation, Li and Sarkar (2010) proposed a method to mask data to protect sensitive information against record linkage disclosure by partitioning a dataset into smaller subgroups to achieve homogeneity in each subgroup. Again, Li and Sarkar (2010) proposed a digression approach that uses the measure for pruning the tree to limit disclosure of sensitive data in the process of building a regression tree model. The

authors also proposed an algorithm that anonymizes both numeric and categorical sensitive data. All three studies were practical in nature. A fourth study by Zuboff (2015) offered more insights on the use and ramifications of big data capitalization and privacy violation in today's digital information era.

Business Value Research on BDPA

Not much has been done to demonstrate how big data analytics can be of value to organizations despite the efforts made by few IS scholars. Chen, Preston, and Swink (2015) adapted the technology–organization– environment (TOE) framework to identify factors that influence the actual usage of big data analytics and how the usage helps with value creation. The authors found that the level of use of big data analytics helps with value creation and is highly influenced by environmental and technological factors that interact with the organization. After investigating the literature for analytics key success factors, Seddon, Constantinidis, Tamm, and Dod (2016) proposed a variance and process model of how analytics contribute to business value. Notwithstanding the capabilities of big data analytics, human know-how still plays a key role in interpreting outcomes, making final decisions and allocating proper resources (Coussement et al., 2015). Hence, there is still a need to understand the human subjective reception and use of analytics knowledge. Accordingly, Shollo and Galliers (2016) introduced a model to delineate how business intelligence systems can impact the knowledge of decision making individuals in their daily practices. Note, that all the studies in this category where published between 2015 and 2016. This shows a need to understand the value of big data analytics within the IS community. We anticipate more publications about this topic in the coming years.

Literature Survey Research on BDPA

Our analysis found 3 studies that reviewed the body of literature on themes vital to BDPA. Shmueli and Koppius (2011) identified six key roles played by predictive analytics in IS research namely; new theory development, measurement development, comparison of competing theories, improvement of existing models, assessment of relevance and assessment of the predictability of an empirical phenomena. Chen et al. (2012) analyzed business intelligence and analytics studies and demonstrated the applications, evolution and research directions of business intelligence and analytics. Hogenboom, Frasincar, Kaymak, de Jong, and Caron (2016) surveyed the literature for techniques to extract information such as textual data from events. Based on our analysis, there is currently no literature survey study on BDPA. The surveys carried out by the above-mentioned studies separately investigate big data and analytics (Chen et al., 2012) or predictive analytics (Shmueli & Koppius, 2011). To the best of our knowledge, our work is the first to systematically investigate the literature on BDPA as one topic.

5. DISCUSSION AND IMPLICATIONS

To advance the IS literature on BDPA, the goal of this paper is to review the literature in order to identify research areas, gaps and applications. We identified a total of 80 studies from top ranked IS journals covering several research topics related to our subject. Our review shows an increase in the number of publications on BDPA, indicating that the use of predictive analytics on big data is growing in importance within the IS community. Nevertheless, at the moment, we noticed that not much has been accomplished in IS research in the area of BDPA. We believe that the main reason the literature is not currently overflown with such studies is that research conducted on BDPA leads to value creation and competitive intelligence for sponsoring organizations that do not want research outcomes published.

State of Current Practice and Implications for Future Research

We classified BDPA studies into five categories namely empirical research, general overview, business value of BDPA, privacy issues of BDPA and literature surveys. Our reflective review reveals several issues that are explained below.

Data Implications of BDPA: On the data aspect, we found that user generated content, particularly through social media and online customer reviews is the most utilized source of data for IS researchers seeking to explore BDPA. Hence there seems to be an underutilization of other wealthy sources of data available to IS researchers. This might be caused by a lack of commitment on the part of some organizations that control ownership of such data. Additionally, our analysis shows that most data used in developing predictive models for health, for instance (Wimmer et al., 2016), were rather small. This may be because of the ethical constraints surrounding patients' data and the fact that some medical conditions for which predictive models were applied are not rampant. Does this mean that small data is actually big data in some scenarios? Furthermore, we found that

studies have analyzed unstructured data that mostly consists of text, web logs, sensor generated data, and images. Finally, we confirm that the complexity of big data introduces important changes in the way information is generated and analyzed (Constantiou & Kallinikos, 2014). For instance, the diversity of big data introduces new challenges for validating data sources. Veracity (i.e., data integrity) was the fourth dimension introduced to label big data. So, with the growing amount of spam accounts created daily on social media and blogs, should organizations trust data form such external sources? It will be interesting for IS studies to provide an understanding of how to validate the authenticity of user generated data and how it affects the outcome of BDPA.

Method Implications of BDPA: In the modeling method aspect of IS research on BDPA we found that longstanding predictive analytics methods have been used in the literature for prediction using big data: decision trees, logistic regression, naïve Bayes, neural networks and support vector machines. This may mean that existing predictive analytics methods can very well handle the complexity of big data. Moreover, we noticed that more recent papers implement the use of advanced algorithms to complement the use to traditional methods particularly if the data is multifarious. For instance, Wasesa, Stam, and van Heck (2017) used a machine learning Gradient boasting method to support regression for better model fitting and prediction. With regards to which predictive technique performs better on big data, we believe that there is no single best method. The supremacy of any predictive modeling method over others is highly dependent on the features of the dataset and goal of the prediction (Soni, 2014). This means that each method can only be suited for certain datasets and problems. The literature calls for a scheme that will help researchers and practitioners in diverse industries to systematically select the most appropriate predictive modeling method to apply to specific big data types and industrial problems. In our review, we also found no consensus on how predictive methods are assessed for BDPA. Therefore, the literature calls for studies to present criteria for evaluating big data predictive models, or studies that can confirm whether or not existing criteria for evaluating predictive models can be applied to big data predictive models.

Industry Application Implications o BDPA: With regards to applications, we found that predictive analytics on big data were applied to several domains, including e-commerce and marketing intelligence, healthcare, financial, security and public safety and utility. Online retailers such as Amazon, Alibaba and Ebay use BDPA to gather insights and thus, predict consumer behavior, improve their CRM (Customer Relationship Management) initiatives, operational efficiency, decision-making and marketing campaigns. Additionally, web mining, text mining, sentiment analysis, opinion mining, and network analysis can be adopted for association rule mining, churn analysis, market basket analysis, campaign analysis, customer life-time value modelling, database segmentation and clustering and anomaly detection for e-commerce and marketing applications (Lim et al., 2013). Insurance companies lose millions to fraudulent claims annually. Using data from previously observed fraudulent patterns, BDPA can be used to detect fraudulent claims and help reduce insurance fraud drastically (Bellini, 2014; Bhattacharyya, Jha, Tharakunnel, & Westland, 2011; Deka, 2014). Insurance companies can use big data to measure potential risk predictors such as demographics, health history and driving records when issuing car or health insurance policies. Banks can utilize data covering credit history, loan applications, and customer data to assess if a customer is likely to default on loan payments. Healthcare practitioners now depend highly on evidence-based medicine from their constantly evolving data to advocate clinical diagnostics for patients, reduce patient waiting times in emergency units and improve managerial operations. BDPA can help improve healthcare outcomes with great efficiency and improved decision-making. We expect that the IS literature will be enriched with healthcare analytics publications in the coming years. The potential of advanced analytics on Security and Public Safety has been delineated by Chen et al. (2012). Law enforcement agencies are utilizing BDPA to improve crime mapping, predicting the likelihood of crime occurrence and anticipating terrorist attacks, among other things (Bachner, 2013; Chen et al., 2012).

Business Value Implications of BDPA: Lately, we noticed an unwavering attention has been dedicated to understanding how big data and analytics can strategically be of value to organizations (Abbasi et al., 2016; Seddon et al., 2016; Shollo & Galliers, 2016). This is a reason why the 5th V (Value) component of big data was recently introduced. Accordingly, IS journals and conferences have introduced specific issues and tracks respectively to call for studies to fill this

void. Shollo and Galliers (2016) reported that organizations face challenges properly implementing business intelligence systems and processing poor quality data in their control. They found that the uniqueness of business intelligence allows for decision makers at different managerial levels to initiate problem articulation and evaluate courses of action to resolve such problems effectively. Seddon et al. (2016) proposed a model that provides a picture-perfect view of an insight-decision-action process of big data analytics and the potential values that could result from the process using Davenport, Harris, and Morison (2010)'s DELTA model of business analytics success factors, Wixom and Watson (2001)'s data warehouse success factors and Seddon, Calvert, and Yang (2010)'s model of organizational benefits from enterprise systems. Although their model has not been fully validated, we expect it to be instrumental in understating the analytics-to-business value chain.

Undoubtedly, the goal of analyzing big data is to provide some sort of value to organizations by leveraging abundant data and related analytics technologies that can help an organization to better understand its business environment (Chen et al., 2012) and guide both future strategies and day-to-day operations (LaValle et al., 2011). We believe that while some studies have justified the practical implications of big data and analytics for organizations (Chen et al., 2015; Seddon et al., 2016; Shollo & Galliers, 2016), there is still more to be accomplished, particularly with regards to the predictive analytics of big data and its value creation, thus we propose several questions in Table 5 (Appendix C).

Privacy Implications of BDPA: It is no secret that we now live in a digital world were every data we generate through our daily activities (e.g., social media, shopping, entertainment) is captured and sold to organizations to better understand our sentiments, opinions and preferences. Big data containing personal data are constantly being used for predictive analytics in several domains, including antiterrorism, crime analysis, marketing research, financial analysis, human behavior study, and healthcare research (Li & Sarkar, 2014). Organizations such as Google have exploited big data as a new profit stream by analyzing datasets and selling user information to organizations. Such practices come with a cost in terms of individual privacy violations of rights and laws (Zuboff, 2015). Collecting data that includes personal information of users requires rigorous ethical authorizations before it is used for research purposes. However, the big data era now makes such data available to organizations to explore without a crystal clear ethical policy in place. Hence, there is a need to understand the ethical and privacy implications of big data and how it can be managed by organizations.

The era of big data and tools to analyze big data has increased the chances of disclosing private and confidential information amongst businesses who share databases either within or across organizations (Menon, Sarkar, & Mukherjee, 2005). Most businesses who share their databases prefer to hide private and confidential information before sharing them even though some of it could be useful to the knowledge discovery from data. Regardless, Li and Sarkar (2006, 2010, 2014) demonstrated that regression trees can be used during analysis to reveal sensitive information about individuals in a dataset even in the realm of existing privacy-preserving methods. This revelation is frightening to say the least. It can expose organizations to lawsuits by their clients who expect their sensitive information to remain confidential. Hence there is much to be accomplished in the area.
To conclude, we propose several research questions from the above discussion Table 5 (Appendix C) to fill the gaps that currently exist in the literature.

Limitations
While this study offers contributions to the understanding of the state of BDPA, it has limitations. First, we observed that studies on BDPA do exist in IS research, but they are not published in premium IS journals. Thus, a limitation is that other suitable BDPA studies may have been excluded from our study. Another limitation is that we did not consider industry published studies in our review. Industrial papers are important in understanding fresh research topics, particularly to help understand business needs, technology types and their applications. In future research, we shall include publications from non-premium IS journals and include industry white papers.

6. CONCLUSION

The complexity of data is growing rapidly as businesses continue to embrace sensors, mobile devices, RFID (Radio frequency identification), audio and video streams, software logs and crowdsourcing systems. Therefore, the use of BDPA has become mainstream for business competition. To examine the present state-of-the-art of BDPA, the current study carries out a

structured review of the academic literature on BDPA. The findings will benefit IS academics who are interested in this nascent concept by offering a comprehensive mapping of research studies in BDPA, and will help practitioners understand the full potential of BDPA.

7. REFERENCES

Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems, 17*, 32.

Abbasi, A., Zahedi, F. M., Zeng, D., Chen, Y., Chen, H., & Nunamaker Jr, J. F. (2015). Enhancing predictive analytics for anti-phishing by exploiting website genre information. *Journal of Management Information Systems, 31*, 109–157.

Abbott, D. (2014). *Applied predictive analytics: principles and techniques for the professional data analyst*: John Wiley & Sons.

Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research, 25*, 443-448. doi: 10.1287/isre.2014.0546

Agrawal, K. (2015). Investigating the determinants of Big Data Analytics (BDA) adoption in Asian emerging economies. *AMCIS 2015 Proceedings*.

Bachner, J. (2013). *Predictive policing: Preventing crime with data and analytics*: IBM Center for the Business of Government.

Baltzan, P., & Welsh, C. (2015). *Business driven information systems* (Fourth Canadian Edition ed.): McGraw-Hill/Irwin.

Banerjee, S., Bhattacharyya, S., & Bose, I. (2017). Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. *Decision Support Systems, 96*, 17.

Bardhan, I., Oh, J.-h., Zheng, Z., & Kirksey, K. (2014). Predictive analytics for readmission of patients with congestive heart failure. *Information Systems Research, 26*, 19–39.

Bauer, J., & Nanopoulos, A. (2014). Recommender systems based on quantitative implicit customer feedback. *Decision Support Systems, 68*, 77–88. doi: 10.1016/j.dss.2014.09.005

Bedeley, R. (2014). BIG DATA OPPORTUNITIES AND CHALLENGES: THE CASE OF BANKING INDUSTRY. *SAIS 2014 Proceedings*.

Bellini, F. (2014). *Big Data Analytics for Financial Frauds Detection*.

Benthaus, J., Risius, M., & Beck, R. (2016). Social media management strategies for organizational impression management and their effect on public perception. *The Journal of Strategic Information Systems, 25*, 127–139. doi: 10.1016/j.jsis.2015.12.001

Beyer, M. A., & Laney, D. (2012). The importance of 'big data': a definition. *Stamford, CT: Gartner*, 2014-2018.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602-613.

Bhimani, A. (2015). Exploring big data's strategic consequences. *Journal of Information Technology, 30*, 66–69. doi: 10.1057/jit.2014.29

Bogaert, M., Ballings, M., & Van den Poel, D. (2016). The added value of Facebook friends data in event attendance prediction. *Decision Support Systems, 82*, 26–34. doi: 10.1016/j.dss.2015.11.003

Breuker, D., Matzner, M., Delfmann, P., & Becker, J. (2016). Comprehensible Predictive Models for Business Processes. *MIS Quarterly, 40*(4), 1009.

Camacho-Collados, M., & Liberatore, F. (2015). A Decision Support System for predictive police patrolling. *Decision Support Systems, 75*, 25-37. doi: 10.1016/j.dss.2015.04.012

Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems, 95*, 91-101. doi: 10.1016/j.dss.2017.01.002

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems, 63*, 67–80. doi: 10.1016/j.dss.2013.08.008

Chen, C. C., Shih, S.-Y., & Lee, M. (2016a). Who should you follow? Combining learning to rank with social influence for informative friend recommendation. *Decision Support Systems*. doi: 10.1016/j.dss.2016.06.017

Chen, D., Preston, D., & Swink, M. (2015). How the Use of Big Data Analytics Affects Value Creation in Supply Chain Management. *Journal of Management Information Systems, 32*, 4-39. doi: 10.1080/07421222.2015.1138364

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly, 36*, 1165–1188.

Chen, L., Li, X., Yang, Y., Kurniawati, H., Sheng, Q. Z., Hu, H.-Y., & Huang, N. (2016b). Personal health indexing based on medical examinations: A data mining approach. *Decision Support Systems, 81*, 54–65. doi: 10.1016/j.dss.2015.10.008

Clarke, R. (2016). Big data, big risks. *Information Systems Journal, 26*, 77–90. doi: 10.1111/isj.12088

Claverie-Berge, I. (2012). Solutions Big Data IBM.

Constantiou, I. D., & Kallinikos, J. (2014). New games, new rules: big data and the changing context of strategy. *Journal of Information Technology, 30*, 44-57. doi: 10.1057/jit.2014.17

Coussement, K., Benoit, D. F., & Antioco, M. (2015). A Bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection. *Decision Support Systems, 79*, 24-32. doi: 10.1016/j.dss.2015.07.006

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems, 80*, 56–71. doi: 10.1016/j.dss.2015.09.003

D'Haen, J., Van den Poel, D., Thorleuchter, D., & Benoit, D. F. (2016). Integrating expert knowledge and multilingual web crawling data in a lead qualification system. *Decision Support Systems, 82*, 69-78.

Dag, A., Topuz, K., Oztekin, A., Bulur, S., & Megahed, F. M. (2016). A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decision Support Systems, 86*, 1–12. doi: 10.1016/j.dss.2016.02.007

Davenport, T. H., Harris, J. G., & Morison, R. (2010). *Analytics at work: Smarter decisions, better results*: Harvard Business Press.

De Cnudde, S., & Martens, D. (2015). Loyal to your city? A data mining analysis of a public service loyalty program. *Decision Support Systems, 73*, 74–84. doi: 10.1016/j.dss.2015.03.004

Deka, G. C. (2014). Big Data Predictive and Prescriptive Analytics. *Handbook of Research on Cloud Infrastructures for Big Data Analytics*, 370.

Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems, 55*(1), 359-363.

Du, X., Ye, Y., Lau, R. Y. K., & Li, Y. (2015). OpinionRings: Inferring and visualizing the opinion tendency of socially connected users. *Decision Support Systems, 75*, 11–24. doi: 10.1016/j.dss.2015.04.007

Farhadloo, M., Patterson, R. A., & Rolland, E. (2016). Modeling customer satisfaction from unstructured data using a Bayesian approach. *Decision Support Systems, 90*, 1.

Fersini, E., Messina, E., & Pozzi, F. A. (2014). Sentiment analysis: Bayesian Ensemble Learning. *Decision Support Systems, 68*, 26–38. doi: 10.1016/j.dss.2014.10.004

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*: Springer.

Franks, B. (2012). *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics* (Vol. 49): John Wiley & Sons.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144.

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using Forum and Search Data for Sales Prediction. *MIS Quarterly, 41*(1), 65.

Goes, P. (2015). *Big Data - Analytics Engine for Digital Transformation: Where is IS?*

Gómez-Vallejo, H. J., Uriel-Latorre, B., Sande-Meijide, M., Villamarín-Bello, B., Pavón, R., Fdez-Riverola, F., & Glez-Peña, D. (2016). A case-based reasoning system for aiding detection and classification of nosocomial infections. *Decision Support Systems, 84*, 104–116. doi: 10.1016/j.dss.2016.02.005

Goul, M., Balkan, S., & Dolk, D. (2015, 2015). *Predictive Analytics Driven Campaign Management Support Systems.* Paper presented at the System Sciences (HICSS), 2015 48th Hawaii International Conference on.

Gualtieri, M., Rowan Curran, A., TaKeaways, K., & To, M. T. B. P. P. (2013). The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013. *Forrester research*.

Hair Jr, J. F. (2007). Knowledge creation in marketing: the role of predictive analytics. *European Business Review, 19*(4), 303-315.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98-115.

Hogenboom, A., Frasincar, F., de Jong, F., & Kaymak, U. (2015). Polarity classification using structure-based vector representations of text. *Decision Support Systems, 74*, 46–56. doi: 10.1016/j.dss.2015.04.002

Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F., & Caron, E. (2016). A Survey of event extraction methods from text for decision support systems. *Decision Support Systems, 85*, 12-22. doi: 10.1016/j.dss.2016.02.006

Huang, T. C.-K., Chen, Y.-L., & Chen, M.-C. (2016). A novel recommendation model with Google similarity. *Decision Support Systems, 89*, 17–27. doi: 10.1016/j.dss.2016.06.005

IBM. (2015). What is big data? Retrieved April 21, 2016, from http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Jiang, P., Liu, X., Zhang, J., & Yuan, X. (2016). A framework based on hidden Markov model with adaptive weighting for microcystin forecasting and early-warning. *Decision Support Systems, 84*, 89–103. doi: 10.1016/j.dss.2016.02.003

Khairul, N. B., & Shahrul, A. M. N. (2015). Efficient identity matching using static pruning q-gram indexing approach. *Decision Support Systems, 73*, 97-108. doi: 10.1016/j.dss.2015.02.015

Kim, J., & Kang, P. (2016). Late payment prediction models for fair allocation of customer contact lists to call center agents. *Decision Support Systems, 85*, 84.

Kiron, D., & Shockley, R. (2011). Creating business value with analytics. *MIT Sloan Management Review, 53*(1), 57.

Koronios, A., Gao, J., & Selle, S. (2014). *BIG DATA PROJECT SUCCESS â€" A META ANALYSIS*.

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*: Morgan Kaufmann.

Kridel, D., & Dolk, D. (2013). Automated self-service modeling: predictive analytics as a service. *Information Systems & e-Business Management, 11*(1), 119-140. doi: 10.1007/s10257-011-0185-1

Langseth, H., & Nielsen, T. D. (2015). Scalable learning of probabilistic latent models for collaborative filtering. *Decision Support Systems, 74*, 1–11. doi: 10.1016/j.dss.2015.03.006

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review, 52*(2), 21-32.

Lee, A. J. T., Yang, F.-C., Chen, C.-H., Wang, C.-S., & Sun, C.-Y. (2016). Mining perceptual maps from consumer reviews. *Decision Support Systems, 82*, 12.

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science, 9*.

Li, H., Wu, J., Liu, L., & Li, Q. (2015). *Adoption of Big Data Analytics in Healthcare: The Efficiency and Privacy*.

Li, X.-B., & Sarkar, S. (2006). Privacy Protection in Data Mining: A Perturbation Approach for Categorical Data. *Information Systems Research, 17*, 254-270. doi: 10.1287/isre.1060.0095

Li, X.-B., & Sarkar, S. (2010). Protecting Privacy Against Record Linkage Disclosure: A Bounded Swapping Approach for Numeric Data. *Information Systems Research, 22*, 774-789. doi: 10.1287/isre.1100.0289

Li, X.-B., & Sarkar, S. (2014). Digression and Value Concatenation to Enable Privacy-Preserving Regression. *Management Information Systems Quarterly, 38*, 679–698.

Li, Y., Thomas, M. A., & Osei-Bryson, K.-M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems*, 1-12. doi: 10.1016/j.dss.2016.07.003

Lim, E.-P., Chen, H., & Chen, G. (2013). Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems, 3*(4). doi: 10.1145/2407740.2407741

Lukoianova, T., & Rubin, V. L. (2014). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online, 24*(1), 4-15.

Luo, X., & Zhang, J. (2013). How Do Consumer Buzz and Traffic in Social Media Marketing Predict the Value of the Firm? *Journal of Management Information Systems, 30*, 213-238. doi: 10.2753/MIS0742-1222300208

Lycett, M. (2013). 'Datafication': Making sense of (big) data in a complex world.

Maciá-Pérez, F., Berna-Martinez, J. V., Fernández Oliva, A., & Abreu Ortega, M. A. (2015). Algorithm for the detection of outliers based on the theory of rough sets. *Decision Support Systems, 75*, 63–75. doi: 10.1016/j.dss.2015.05.002

Markus, M. L. (2015). New games, new rules, new scoreboards: the potential consequences of big data. *Journal of Information Technology, 30*, 58–59. doi: 10.1057/jit.2014.28

Martens, D., & Provost, F. (2014). Explaining Data-Driven Document Classifications. *Management Information Systems Quarterly, 38*, 73–99.

Martens, D., Provost, F., Clark, J., & de Fortuny, E. J. (2016). Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics. *MIS Quarterly, 40*(4), 869.

Meire, M., Ballings, M., & Van den Poel, D. (2016). The added value of auxiliary data in sentiment analysis of Facebook posts. *Decision Support Systems, 89*, 98–112. doi: 10.1016/j.dss.2016.06.013

Menon, S., & Sarkar, S. (2016). Privacy and Big Data: Scalable Approaches to Sanitize Large Transactional Databases for Sharing. *MIS Quarterly, 40*(4), 963.

Menon, S., Sarkar, S., & Mukherjee, S. (2005). Maximizing Accuracy of Shared Databases when Concealing Sensitive Patterns. *Information Systems Research, 16*, 256-270. doi: 10.1287/isre.1050.0056

Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A Machine Learning Approach to Improving Dynamic Decision Making. *Information Systems Research, 25*, 239-263. doi: 10.1287/isre.2014.0513

Mishra, R., Kumar, P., & Bhasker, B. (2015). A web recommendation system considering sequential information. *Decision Support Systems, 75*, 1-10. doi: 10.1016/j.dss.2015.04.004

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems, 72*, 72-81. doi: 10.1016/j.dss.2015.02.007

Müller, O., Junglas, I., Brocke, J. v., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems, 25*, 289-302. doi: 10.1057/ejis.2016.2

Nettleton, D. (2014). *Commercial Data Mining: Processing, Analysis and Modeling for Predictive Analytics Projects*: Elsevier.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems, 24*, 3–14. doi: 10.1016/j.jsis.2015.02.001

Pai, H.-T., Wu, F., & Hsueh, P.-Y. S. S. (2014). A relative patterns discovery for enhancing outlier detection in categorical data. *Decision Support Systems, 67*, 90–99. doi: 10.1016/j.dss.2014.08.006

Peffers, K., & Ya, T. (2003). Identifying and evaluating the universe of outlets for information systems research: Ranking the journals. *Journal of Information Technology Theory and Application (JITTA), 5*(1), 6.

Phillips-Wren, G., Iyer, L. S., Kulkarni, U., & Ariyachandra, T. (2015). Business Analytics in the Context of Big Data: A Roadmap for Research. *Communications of the Association for Information Systems, 37*, 23.

Pournarakis, D. E., Sotiropoulos, D. N., & Giaglis, G. M. (2017). A computational model for mining consumer perceptions in social media. *Decision Support Systems, 93*, 98-105,108-110. doi: 10.1016/j.dss.2016.09.018

Power, D. (2013). *Using Big Data for Analytics and Decision Support*.

Russell, J. (2015). Predictive analytics and child protection: Constraints and opportunities. *Child abuse & neglect, 46*, 182-189.

Saboo, A. R., Kumar, V., & Park, I. (2016). Using Big Data to Model Time-Varying Effects for Marketing Resource (Re)Allocation. *MIS Quarterly, 40*(4), 911.

Sahoo, N., Krishnan, R., Duncan, G., & Callan, J. (2011). Research Note—The Halo Effect in Multicomponent Ratings and Its Implications for Recommender Systems: The Case of Yahoo! Movies. *Information Systems Research, 23*, 231-246. doi: 10.1287/isre.1100.0336

Sahoo, N., singh, p. v., & Mukhopadhyay, T. (2012). A Hidden Markov Model for Collaborative Filtering. *Management Information Systems Quarterly, 36*, 1329–1356.

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. doi: 10.1016/j.dss.2015.10.006

Schumaker, R. P., Jarmoszko, A. T., & Labedz, C. S., Jr. (2016). Predicting wins and spread in the Premier League using a sentiment analysis of twitter. *Decision Support Systems, 88*, 76.

Seddon, P. B., Calvert, C., & Yang, S. (2010). A multi-project model of key factors affecting organizational benefits from enterprise systems. *MIS Quarterly, 34*(2), 305-328.

Seddon, P. B., Constantinidis, D., Tamm, T., & Dod, H. (2016). How does business analytics contribute to business value? *Information Systems Journal*, n/a-n/a. doi: 10.1111/isj.12101

Sharma, R., Mithas, S., & Kankanhalli, A. (2014). Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations. *European Journal of Information Systems, 23*, 433-441. doi: 10.1057/ejis.2014.17

Shim, J. P., French, A. M., Guo, C., & Jablonski, J. (2015). *Big Data and Analytics: Issues, Solutions, and ROI* (Vol. 37).

Shmueli, G., & Koppius, O. (2011). Predictive analytics in information systems research. *Management Information Systems Quarterly, 35*(3), 553-572.

Shollo, A., & Galliers, R. D. (2016). Towards an understanding of the role of business intelligence systems in organisational knowing. *Information Systems Journal, 26*, 339-367. doi: 10.1111/isj.12071

Shynkevich, Y., McGinnity, T. M., Coleman, S. A., & Belatreche, A. (2016). Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning. *Decision Support Systems, 85*, 74.

Soni, S. (2014). Overview of Predictive Modeling Approaches in Health Care Data Mining. *Emerging Methods in Predictive Analytics: Risk Management and Decision-Making: Risk Management and Decision-Making*, 349.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems, 29*, 217-248. doi: 10.2753/MIS0742-1222290408

Sun, Z., Zou, H., & Strang, K. (2015). Big Data Analytics as a Service for Business Intelligence. In M. Janssen, M. Mäntymäki, J. Hidders, B. Klievink, W. Lamersdorf, B. van Loenen, & A. Zuiderwijk (Eds.), *Open and Big Data Management and Innovation : 14th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2015, Delft, The Netherlands, October 13-15, 2015, Proceedings* (pp. 200-211). Cham: Springer International Publishing.

Tsai, C.-F., & Chen, Z.-Y. (2014). Towards high dimensional instance selection: An evolutionary approach. *Decision Support Systems, 61*, 79-92. doi: 10.1016/j.dss.2014.01.012

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems, 75*, 38–48. doi: 10.1016/j.dss.2015.04.013

Visinescu, L. L., & Evangelopoulos, N. (2014). Orthogonal rotations in latent semantic analysis: An empirical study. *Decision Support Systems, 62*, 131–143. doi: 10.1016/j.dss.2014.03.010

Volkov, A., Benoit, D. F., & Van den Poel, D. (2017). Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination. *Decision Support Systems, 98*, 59.

Wang, W., Zhang, G., & Lu, J. (2016). Member contribution-based group recommender system. *Decision Support Systems, 87*, 80.

Wasesa, M., Stam, A., & van Heck, E. (2017). The seaport service rate prediction system: Using drayage truck trajectory data to predict seaport service rates. *Decision Support Systems, 95*, 37-48. doi: 10.1016/j.dss.2016.11.00

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems, 34*(1), 1247-1268.

Wattal, S., Telang, R., Mukhopadhyay, T., & Boatwright, P. (2011). What's in a "Name"? Impact of Use of Customer Information in E-Mail Advertisements. *Information Systems Research, 23*, 679-697. doi: 10.1287/isre.1110.0384

Wimmer, H., Yoon, V. Y., & Sugumaran, V. (2016). A multi-agent system to support evidence based medicine and clinical decision making via data sharing and data privacy. *Decision Support Systems, 88*, 51–66. doi: 10.1016/j.dss.2016.05.008

Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision Support Systems*. doi: 10.1016/j.dss.2016.06.016

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., . . . Turetken, O. (2014). The current state of business intelligence in academia: The arrival of big data. *Communications of the Association for Information Systems, 34*(1), 1.

Wixom, B. H., & Watson, H. J. (2001). An empirical investigation of the factors affecting data warehousing success. *MIS Quarterly*, 17-41.

Woerner, S. L., & Wixom, B. H. (2015). Big data: extending the business strategy toolbox. *Journal of Information Technology, 30*, 60–62. doi: 10.1057/jit.2014.31

Wu, F., Huang, Y., Song, Y., & Liu, S. (2016). Towards building a high-quality microblog-specific Chinese sentiment lexicon. *Decision Support Systems, 87*, 39–49. doi: 10.1016/j.dss.2016.04.007

Yoo, Y. (2015). It is not about size: a further thought on big data. *Journal of Information Technology, 30*, 63–65. doi: 10.1057/jit.2014.30

Yuan, H., Lau, R. Y. K., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*. doi: 10.1016/j.dss.2016.08.001

Zeng, D. (2015). Crystal Balls, Statistics, Big Data, and Psychohistory: Predictive Analytics and Beyond. *IEEE Intelligent Systems*(2), 2-4.

Zhang, M., Guo, X., & Chen, G. (2016). Prediction uncertainty in collaborative filtering: Enhancing personalized online product ranking. *Decision Support Systems, 83*, 10.

Zheng, B., Zhang, J., Yoon, S. W., Lam, S. S., Khasawneh, M., & Poranki, S. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. *Expert Systems with Applications, 42*(20), 7110-7120.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*, 75-89. doi: 10.1057/jit.2015.5

APPENDIX A

	Author	Method	Data Type	Key Contributions	Application
P1	Li and Sarkar (2006)	N/A	Multiple sources	A method for limiting exposure of confidential information during data classification task	N/A
P2	Li and Sarkar (2010)	N/A	Multiple sources	A data-masking method for protecting private information against record linkage disclosure	N/A
P3	Shmueli and Koppius (2011a)	N/A	N/A	Highlighted the six roles of predictive analysis and conducted a literature review on the topic.	N/A
P4	Sahoo, Krishnan, Duncan, and Callan (2011)	Flexible mixture model and EM algorithm (Bayesian network)	Movie rating data	Improving collaborative filtering recommendation using multiple component rating	Recommender Systems
P5	Sahoo et al. (2012)	Hidden Markov Model	Social media blog data	A hidden markov model for making personalized recommendations when with changing user preferences overtime.	Recommender systems
P6	Wattal et al. (2011)	Hierarchical Bayesian model	Emails messages	A customer segmentation model for target email advertisement	Customer segmentation
P7	Chen et al. (2012)	N/A	N/A	Discussed the evolution, applications and research directions in business intelligence and analytics	N/A
P8	Stieglitz and Dang-Xuan (2013)	SentiStrength and Regression Analysis	Twitter post	Sentiment analysis of social media post and user's information sharing behavior	Sentiment analysis
P9	Luo and Zhang (2013)	Regression Analysis VARX models	Consumer reviews	A model for predicting firm value through customer attitude of blogs, social media posts, user generated reviews e.t.c	Financial
P10	Meyer et al. (2014)	Decision tree: C4.5 algorithm	Patient healthcare data	A model for optimal performance of dynamic decision-making strategies	Healthcare And Manufacturing
P11	Pai et al. (2014)	Unsupervised Approach	Multiples sources (8)	An unsupervised method for outlier detection	Fraud/Anomaly Detection
P12	Li and Sarkar (2014)	Regression trees	Census data set	A pruning method to limit exposure of sensitive data A method to anonymize data during knowledge discovery	Generic
P13	Agarwal and Dhar (2014)	N/A	N/A	Editorial	N/A
P14	Martens and Provost (2014)	An SEDC search algorithm	Website content	A search algorithm to deal with high dimensional data	Text mining
P15	Constantiou and Kallinikos (2014)	N/A	N/A	Discussed the strategic implication of big data	N/A
P16	Visinescu and Evangelopoulos (2014)	Factor analysis	-Text massages -DHS Idea data -news report	Compared three types of orthogonal rotations (Varimax, Quartimax and Equamax	Latent semantic analysis
P17	Bardhan et al. (2014)	Regression model (logit model)	Patient admission records	A model to predict remission within 30 days of discharge	Healthcare
P18	Bauer and Nanopoulos (2014)	Matrix factorization	Multiple product review data sets	Proposed a new algorithm for recommendation based on quantitative implicit customer feedback using matrix factorization	Recommender Systems
P19	Fersini et al. (2014)	Bayesian Ensemble Learning approach	Movie reviews	A novel ensemble approach for sentiment classification purposes	Sentiment analysis
P20	Tsai and Chen (2014)	Efficient Genetic algorithm	Multiple sources: documents	A method for reducing high dimensional data for classification purpose	Generic
P21	Sharma et al. (2014)	N/A	N/A	Editorial	N/A
P22	Chang et al. (2014)	N/A	N/A	Discussed the philosophical change introduced by big data "Theory no longer matters"	N/A
P23	Coussement et al. (2015)	Bayesian approach	Online customer reviews	A Bayesian decision support system framework that integrates human	Customer satisfaction detection

				expert subjective opinion with organizational data information	
P24	Camacho-Collados and Liberatore (2015)	Time series models	Theft reports	A decision support system for Crime Prediction	Security and Public Safety
P25	Maciá-Pérez, Berna-Martinez, Fernández Oliva, and Abreu Ortega (2015)	Rough Sets Theory	Census data	An outlier detection algorithm using rough sets theory	Fraud/ Anomaly detection
P26	Van Vlasselaer et al. (2015)	Random forest	Credit card transactions	A fraud detection system for credit card transactions	Fraud/ Anomaly detection
P27	Mishra, Kumar, and Bhasker (2015)	Rough set based similarity	Website content	A web recommender system that is based on the rough set similarity theory to allow for overlapping clusters	Recommender systems
P28	Zuboff (2015)	N/A	N/A	Discussed privacy implications of big data	N/A
P29	Khairul and Shahrul (2015)	Q-gram indexing	Census data set	A method for identity matching in large datasets	Identity matching
P30	Abbasi et al. (2015)	Tree Kernel	Website content	A method for detecting phishing websites	Fraud/ Anomaly detection
P31	Bhimani (2015)	N/A	N/A	Commentary	N/A
P32	Cresci et al. (2015)	Decision tree	Tweeter Accounts	An algorithm to detect fake twitter followers	Fraud/ Anomaly detection
P33	Chen et al. (2015)	N/A	N/A	Modeled the factors that influence the use of big data analytics and the organizational outcomes of the use of big data analytics.	N/A
P34	Moeyersoms and Martens (2015)	SVM	Customer records	A method for modeling by including high-cardinality attributes	Energy Sector: Churn prediction
P35	Yoo (2015)	N/A	N/A	Commentary	N/A
P36	De Cnudde and Martens (2015)	Naïve Bayes	Transaction data	A model for customer loyalty programs in public service	Public Service
P37	Markus (2015)	N/A	N/A	Commentary	N/A
P38	Du, Ye, Lau, and Li (2015)	The Weighted-vote Relational Neighbor with Relaxation Labeling (wvRNRL) algorithm	Political blogs data	An algorithm to extract and visualize social intelligence from social media to support decision making	Opinion mining and prediction
P39	Hogenboom, Frasincar, de Jong, and Kaymak (2015)	SVM	Movie review data	A sentiment analysis feature extraction framework	Sentiment analysis
P40	Langseth and Nielsen (2015)	Probabilistic collaborative filtering model based on Bayes framework	Movie rating data	Proposed a scalable learning scheme for a probabilistic generative model for collaborative filtering	Recommender Systems
P41	Newell and Marabelli (2015)	N/A	N/A	Discussed privacy implications of big data	N/A
P42	Gómez-Vallejo et al. (2016)	Decision Tree	Patient health data	A case-based reasoning (CRB) system for detecting and classifying Nosocomial infections	Healthcare
P43	Jiang et al. (2016)	Continuous Hidden Markov Model; Adaptive Exponential smoothing and PCA	Lake data	A framework for forecasting microsystin	Biological and Diseases Control
P44	Wimmer et al. (2016)	Naïve Bayes	UCI Brest cancer data set	A multi agent framework that facilitates data sharing and integration for evidence based medicine	Healthcare
P45	Huang et al. (2016)	Googlle similarity	User movie ratings	An item-based collaborative filtering systems using rating matrix	Recommender systems
P46	Dag et al. (2016)	Decision trees (C&RT)	UNOS heart transplantation	A survival prediction model for heart transplantation	Healthcare

P47	Li, Thomas, and Osei-Bryson (2016)	N/A	N/A	A model for knowledge discovery through big data analytics	N/A
P48	Hogenboom et al. (2016)	N/A	N/A	Reviewed several data-driven, knowledge-driven and hybrid methods for event extraction.	N/A
P49	Clarke (2016)	N/A	N/A	Discussed the problems and opportunities of big data	N/A
P50	Abbasi et al. (2016)	N/A	N/A	Editorial: Discussed the theoretical and methodological opportunities, challenges and implications of big data	N/A
P51	Seddon et al. (2016)	N/A	N/A	A variance and process model or how business analytics contributes to business value	N/A
P52	Chen et al. (2016b)	MyPHI	Geriatric medical examination records	Proposed MyPHI, a data mining method for predicting personal health index	Healthcare
P53	Salehan and Kim (2016)	SentiStrength	Product reviews	A predictive model for the performance of customer reviews in terms of readership and helpfulness	Sentiment analysis
P54	Benthaus, Risius, and Beck (2016)	A mixed method approach (The SentiStrength algorithm; Naïve Bayesian filter)	Microblogging from twitter data	Evaluated the effect of different social media strategies on perception of the public	Sentiment analysis
P55	Meire, Ballings, and Van den Poel (2016)	Random forest	Facebook post	A sentiment prediction model, including leading information, lagging information, and traditional post variables	Sentiment analysis
P56	Bogaert et al. (2016)	Adaboost	Facebook post	Evaluates friends network data to predict event attendance	Event attendance prediction
P57	Yuan et al. (2016)	Random forest	News articles	Proposed a text analytics framework for crowdfunding analysis	Financial: Crowdfunding
P58	Shollo and Galliers (2016)	N/A	N/A	Evaluates the performance outcomes of business intelligence systems in organizational knowing	N/A
P59	Wu et al. (2016)	Sentiment lexicon	Microblog messages	A microblog-specific Chinese sentiment lexicon.	Sentiment analysis
P60	Winkler et al. (2016)	Smoke list	Product reviews	A method for discovering danger words indicative of toy safety defects is proposed.	Toy safety defection through danger world list
P61	Chen et al. (2016a)	Modified matrix factorization	User review and user generated content	A method for using social network for friend recommendation	Recommender Systems
P62	Farhadloo, Patterson, and Rolland (2016)	Bayesian network/Models	User reviews and ratings	A model to predict customer satisfaction	Sentiment Analysis
P63	Meire et al. (2016)	Random Forest	Social media	A model to determine key predictors and relationships to sentiment outcomes	Sentiment Analysis
P64	Schumaker, Jarmoszko, and Labedz (2016)	Other Methods	Social media	A model to predict soccer game outcome using pre-game tweets in social media.	Sentiment Analysis
P65	Wang, Zhang, and Lu (2016)	Other Methods	User reviews and ratings	A model for group profiles recommendation by considering all membership contributions to groups activities	Recommender Systems
P66	Shynkevich et al. (2016)	Multiple kernel learning	Text Documents	A model for stock price prediction	Stock market prediction
P67	Kim and Kang (2016)	Hybrid method	Customer data	A prediction model for scoring and collecting debts	Call centers
P68	Zhang et al. (2016)	Other Methods	User reviews and ratings	A method for predicting uncertainty	Recommender Systems
P69	D'Haen et al. (2016)	Regression	Website content	Proposed a method for using web data as input for decision support systems	Market Intelligence: Sales
P70	Lee, Yang, Chen, Wang, and Sun (2016)	Mining perceptual maps (MPM)	User reviews and ratings	Proposed a method to build perceptual maps and radar charts from large datasets.	Sentiment and Opinion Analysis

P71	Martens, Provost, Clark, and de Fortuny (2016)	Hybrid method	Customer data	Evaluates the use of massive fine-grained data for target marketing using customer behavioral patterns	Market Intelligence: Target Marketing
P72	Menon and Sarkar (2016)	Other Methods	Customer data	Proposed an approach that hides sensitive information during predictive analytics to promote data sharing	Recommender Systems
P73	Saboo, Kumar, and Park (2016)	Time series	Customer data	A predictive model that reveals changes in the effect of marketing programs overtime	Market Intelligence: Resource Allocation for target marketing
P74	Breuker, Matzner, Delfmann, and Becker (2016)	RegPFA	Customer data	Introduced a predictive modelling approach for business process event data	Business Process Mining
P75	Volkov, Benoit, and Van den Poel (2017)	Markov for discrimination and Random forest	Financial data	A predictive model for financial bankruptcy measured using different datasets from multiple time periods	Financial Applications
P76	Banerjee et al. (2017)	Regression	User reviews and ratings	Evaluated the impact of review trustworthiness and the moderating relationship between review-based online reputation and business patronage.	Recommender Systems
P77	Carneiro et al. (2017)	Random Forest	Customer data	A risk scoring model for fraud detection	Fraud Detection
P78	Wasesa et al. (2017)	Regression with GBM	Others	Service rate prediction system to optimize truck pick-up/delivery operations at seaports	Logistics
P79	Pournarakis et al. (2017)	Generic algorithms	Social media	A computational model that mine influential topics and customer perception using social media data to improve target marketing	Market Intelligence: Target Marketing
P80	Geva et al. (2017)	Regression	Multiple sources: Sales data, Search engine data logs, social media data	Explored the relationships between to data sources (social media data and search engine logs) and their impact of sales prediction outcomes	Sales prediction

Table 3: Summary of selected BDPA papers

APPENDIX B

Big Data Characteristics	Count	%	Analysis Techniques	Count	%
1 V	15	25.0%	Other Methods	23	38%
2Vs	31	51.7%	Regression method	9	15%
3Vs	14	23.3%	Bayesian network/Models	5	8%
Total	**60**	**100%**	Random forest	5	8%
			Decision Tree	4	7%
Data Size	**Count**	**%**	Innovative Algorithms	3	5%
Less than a 10,000	9	15%	Support Vector Machines	3	5%
10,000 to 100,000	19	32%	Matrix factorization	2	3%
100, 00 to 1,000,000	16	27%	Rough Set	2	3%
1,000,000 to 10,000,000	7	12%	Naïve Bayes	2	3%
10,000,000 and above	9	15%	Time series	2	3%
Total	**60**	**100%**	**Total**	**60**	**100%**
Data Source	**Count**	**%**	**Application Domain**	**Count**	**%**
User reviews and ratings	16	27%	Sentiment Analysis	11	18%
Social media	10	17%	Recommender Systems	11	18%
Transaction records	9	15%	Others	10	17%
Health records	7	12%	Market Intelligence	9	15%
Others	6	10%	Healthcare	6	10%
Website content	4	7%	Anomaly/Fraud detection	6	10%
Census data	3	5%	Financial Applications	4	7%
Text Documents	3	5%	Security and Public Safety	2	3%
Email/Text messages	2	3%	Text Mining	1	2%
Total	**60**	100%	**Total**	**60**	**100%**

Table 4: Summary of Review Findings

APPENDIX C

Research Topics		Potential Research Questions
Empirical Research	Data level	• How do the dimensions (i.e., volume, variety, velocity, veracity and value) of big data affect predictive analytics outcome? • How much data (i.e., volume, variety and velocity) is qualified as big data for predictive analytics?
	Method Level	• How are traditional predictive analytics methods used to analyze big data? • How can we determine what predictive analytics methods is suitable for specific big data problems? • What are the guidelines for evaluating big data predictive models? • How can we analyze unstructured big data (i.e., images, videos, sounds, etc.) for prediction? • What are the visualization methods for BDPA?
	Application Level	• What are the implications of BDPA for inside sales?
Business Value of BDPA		• How can organizations achieve better performance and gain competitive advantage from BDPA? • How can organizations measure the value of BDPA? • What are the business risks of using BDPA? And, how can the risks be effectively managed. • Are investments in BDPA strategically sustainable? • What are the implementation strategies for BDPA?
Ethics and Privacy		• What are the ethical and privacy implications of big data? • How can big data ethical and privacy issues be effectively managed?

Table 5: Future Research Opportunities