Elragal and Klischewski *J Big Data* (2017) 4:19
DOI 10.1186/s40537-017-0079-2

Journal of Big Data

RESEARCH Open Access

Theory-driven or process-driven prediction? Epistemological challenges of big data analytics

Ahmed Elragal[1*] and Ralf Klischewski[2]

*Correspondence:
ahmed.elragal@ltu.se
[1] Department of Computer Science, Electrical and Space Engineering, Computer and Systems Science, Luleå University of Technology, 971 87 Luleå, Sweden
Full list of author information is available at the end of the article

Abstract

Most scientists are accustomed to make predictions based on consolidated and accepted theories pertaining to the domain of prediction. However, nowadays big data analytics (BDA) is able to deliver predictions based on executing a sequence of data processing while seemingly abstaining from being theoretically informed about the subject matter. This paper discusses how to deal with the shift from theory-driven to process-driven prediction through analyzing the BDA steps and identifying the epistemological challenges and various needs of theoretically informing BDA throughout data acquisition, preprocessing, analysis, and interpretation. We suggest a theory-driven guidance for the BDA process including acquisition, pre-processing, analytics and interpretation. That is, we propose—in association with these BDA process steps—a lightweight theory-driven approach in order to safeguard the analytics process from epistemological pitfalls. This study may serve as a guideline for researchers and practitioners to consider while conducting future big data analytics.

Keywords: Big data analytics, Epistemological challenges, Information systems theories, Predictive research

Background

Scientists are accustomed to make predictions based on consolidated and accepted theories pertaining to the domain of prediction. However, nowadays big data analytics (BDA) is able to deliver predictions based on executing a sequence of processing while seemingly abstaining from being theoretically informed about the subject matter. Seizing these new opportunities is tempting: some researchers have been trapped by the sheer amount of datasets made available by leading data-driven companies, which are either directed towards the companies' own prosperity or representing rather small subsets (e.g. of users). For example, understanding the differences between the vast majority of users (i.e. humanity) and smaller subsets of people, whose activities are captured in big datasets, is critical to correct analysis of the data. Surely, BDA needs exploration, but at the same time also reflection to guide BDA research to a prospering future.

Big data analytics research applies machine learning, data mining, statistics, and visualization techniques in order to collect, process, analyze, visualize, and interpret results [1]. BDA, as a process, is based on many disciplines that analyze data to elucidate hidden

knowledge. Yet, BDA research either employs exploratory data analysis to generate hypotheses, or alternatively pursues predictions relying heavily on advanced machine learning, data mining and statistical algorithms.

Our analysis and argument focuses on predictive research since it lends itself to BDA more than exploratory research. We share the point of view (see e.g. [2]) that for BDA to be useful in the long run, it needs epistemological reflection and it needs also to be theory-driven, not only driven by data that is easily available. However, the question addressed here is: *How to address the epistemological challenges in the process of BDA?* Accordingly, we analyze the sequence of processing in BDA and seek to identify the various needs of theoretically informing BDA throughout all of its steps.

This paper is organized as follows. The next section briefly reviews the role of scientific theory in generating predictions. Then we clarify the term BDA and discuss various challenges that already have been identified. The core of the paper analyzes BDA as a process: data acquisition, preprocessing, analysis, interpretation—and for each step we examine and exemplify required critical decisions and point to the underlying epistemological problems and possible solutions based on adequate theoretical foundation. Before concluding we discuss (a) the theoretical groundwork to be done in order to lead BDA to safe and prosperous grounds as well as (b) how to deal with the shift from theory to process in BDA-based prediction.

Science, theory, and predictability

The philosophy of science is rich of different schools of thought. From ancient times until today the ontological and epistemological underpinnings have changed and depend on the tradition and interest of the researcher. For example, science can be discovery (in the school of positivism) or social construction, and accordingly the function of theory and their role for scientific prediction vary widely. However, even the harshest critics of universally fixed understanding of science, e.g. [3] who advanced the idea of epistemological anarchism, agree that science entails a disciplined way to study the natural and/or socially constructed world. In that line the word 'science' has become increasingly associated with the scientific method itself, i.e. the way how scientist interrelate the 'facts', i.e. the empirical data which they are able to constitute, and the 'theory', which supposedly captures the scientific knowledge for reuse such as explanation and/or prediction. However, the production of scientific knowledge has always been facing plenty of epistemological challenges because a universal basis for 'how' to acquire knowledge has never emerged and therefore every (new) approach remains subject to criticism from various perspectives.

The debate about scientific theories and the call for pluralism is more vivid in the social sciences (compared to natural and formal sciences) where observation and data collection is much more depending on the worldview of the researcher. Building on previous works [4] defines "theory as a statement of relationships between units observed or approximated in the empirical world", where 'observed' means measureable and 'approximated' means constructed whenever which the very nature of the unit of study cannot be observed directly (e.g., centralization, satisfaction, or culture). The primary goal of a theory is to answer questions of knowledge seekers: not only *what* (descriptive), but also *how*, *when*, and *why* [4, 5].

The utility (if not quality) of theories is usually considered as a function of (a) overall explanatory power, i.e. the ability to "describe and explain a process or sequence of events", and (b) the predictive power, i.e. to "understand and predict outcomes of interest, even if only probabilistically" [6]; with reference to previous works. In social science, the importance of theory prediction is understood as being related to the sample size: "Given a large enough sample, and/or a long enough period of observation, theorists can predict on the basis of some of the worst explanations or no explanations at all. In other words, given a large enough sample and/or a long enough period of observation, one is able to predict for all the wrong reasons" [4], p. 509f. For example, the prediction that a tossed coin will land heads up half of the time is accurate just because of statistics (if the coin is tossed in the air often enough), not because of any domain-related theory. On the contrary, theory-based prediction implies an understood cause–effect relation that, for example, predicts job satisfaction caused by job enrichment and participatory decision making, but only for a limited scope of organizations and employees.

The distinction of explanatory power and predictive power is also known in information systems (IS) research which is an interdisciplinary endeavour between computer science and social/management science, trying to understand and support socio-technical systems. For example, [7] distinguishes five different types of theories: for analyzing and describing, for understanding, for predicting, for explaining and predicting, and for design and action. In this context we focus on the difference between explanatory models that aim to statistically test theory-driven hypotheses using empirical data (according to [8] still dominating the IS literature) and predictive models that aim to make predictions based on models. Predictive studies include inductive discovery of relationships among variables in a given dataset, whereby the discovery is driven by techniques and algorithms, without testable a priori hypotheses about causal relationships to be explicitly formulated (e.g. [9]).

Nowadays many practical examples illustrate this shift: Google's language translator does not 'understand' language, nor do its algorithms know the contents of webpages. IBM's Watson does not understand the questions it is asked or use deep causal knowledge to generate questions to the answers it is given. There are dozens of lesser-known companies that likewise are able to predict the odds of someone responding to a display ad without a solid theory but rather based on chunks of data about the behavior of individuals and the similarities and differences in that behavior [1].

With the availability of an abundance of data and computing power to process this data, it seems as if the strive for probabilistic predictability will take over, and scientific utility can be achieved through data processing with less or even without theory. It seems as if the fruitful and seemingly inevitable separation of inductive and deductive research is challenged by data science as a 'competitive' approach, i.e. to extract knowledge or insights from data without a priori theories and without theoretical reflection. But is data science, and data analytics in particular, indeed a scientific method free of theoretical input?

Kitchin [2] analyzes that epistemologically BDA is tempted to fall into the traps of empiricism (with bias in sampling, interpretation, etc.) and rather advocates data-driven science as "a reconfigured version of the traditional scientific method, providing a new way in which to build theory" (p. 6). It combines different approaches that are abductive

(neglecting irrelevant data relations), inductive (generating propositions) and deductive (testing propositions). All of these approaches deal with theories, yet not as starting or end points, but focused on and related to the steps of the data analysis process.

If we are to expect that theory building and predictability are increasingly an outcome of (big) data processing instead of a reflected cycle of inductive and deductive research, then indeed we have to reassess the epistemological underpinnings of our research process. We aim to contribute to this discussion by focusing on the epistemological challenges of every step in the BDA process and seek to point out theoretical development in order to support BDA in the future.

Big data analytics

It is difficult nowadays to open a popular publication today, online or in the physical world and not run into a reference to data science, analytics, big data, or some combination thereof [10]. Big data are data whose scale, distribution, diversity, and velocity require the use of technical architectures, analytics, and tools in order to enable insights that reveal hidden knowledge and create value to business. Three main features characterize big data: volume, variety, and velocity (aka the three V's). The volume of the data is its size, and how enormous it is. Velocity refers to the rate with which data is changing, or how often it is created. Finally, variety includes the different formats and types of data, as well as the different kinds of uses and ways of analyzing the data. Data volume is the primary attribute of big data [11]. Big data can be quantified by size in TBs or PBs, as well as even the number of records, transactions, tables, or files. Additionally, one of the things that make big data really big is that it is coming from a greater variety of sources than ever before, including IoT data, logs, clickstreams, and social media. Using these sources for analytics means that common structured data is now joined by unstructured data, such as text and human language, and semi-structured data, such as extensible markup language (XML), JSON or rich site summary (RSS) feeds. Furthermore, multi-dimensional data can be drawn from a data warehouse to add historic context to big data. Thus, with big data, variety is just as big as volume. Moreover, big data can be described by its velocity or speed. This is basically the frequency of data generation or the frequency of data delivery. The leading edge of big data is streaming data, which is collected in real-time from the websites. Some researchers and organizations have discussed the addition of a fourth V, or veracity. Veracity focuses on the quality of the data. This characterizes big data quality as good, bad, or undefined due to data inconsistency, incompleteness, ambiguity, latency, deception, and approximations [12].

The interest in BDA research is on the increase. Google's adoption of the MapReduce was definitely a catalyst, which has led to a lot of developments in the area of BDA. Further, the development and deployment of Apache Hadoop, SPARK, and Mahout has also opened the doors for organizations to process extremely large datasets that has never been possible. BDA is the use of advanced techniques, mostly data mining and statistical, to find (hidden) patterns in (big) data. BDA is where advanced techniques operate on big datasets [13]. The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems [12]. A significant amount of these techniques rely on commercial tools such as relational DBMS, data warehousing, ETL, OLAP, and business analytics tools. During

the IEEE 2006 International Conference on Data Mining (ICDM), the top-ten data mining algorithms were defined based on expert nominations, citation counts, and a community survey. In order, those algorithms are: C4.5, k-means, support vector machine (SVM), Apriori, expectation maximization (EM), PageRank, AdaBoost, k-nearest neighbors (kNN), Naïve Bayes, and CART. They cover classification, clustering, regression, association analysis, and network analysis. Actually, not only organizations and governments generate data; each and every one of us now is a data generator [14]. We produce data using our mobile phones, social networks interactions, GPS, etc. Most of such data, however, is not structured in a way so as to be stored and/or processed in traditional DBMS. This calls for BDA techniques in order to make sense out of such data.

Big data analytics is inherently related to data mining, a term that has often been used interchangeably with knowledge discovery in database (KDD). However, we see data mining as a step towards knowledge discovery. The term KDD was coined in 1989 to point to the process of finding knowledge in data [15]. KDD is also defined as the process of finding patterns hidden information or unknown facts in the database. Traditionally the notion of finding useful unknown patterns and hidden information in raw data has been given many titles including knowledge discovery in database, data mining, data archaeology, information discovery, knowledge discovery or extraction, and information harvesting. The lack of consensus on the term is attributable to the relative novelty as well as the multi-disciplinary nature of KDD. Multi-disciplinary means that KDD belongs to many disciplines like statistics and computer [machine learning, artificial intelligence (AI), databases, data warehousing, expert systems, knowledge acquisition and data visualization]. Data mining is considered a step in the KDD process of discovering useful knowledge from data while data mining points to the application algorithm or technique used for extracting patters and unknown information from the raw data.

Big data analytics is mostly used with the intention to predict. Prediction is the ability to foresee the future, based on applying certain techniques on datasets. Predictive analytics is a process whereby information extracted from various data sources is utilized to elucidate patterns as well as predict the future. Predictive analytics has the potentials to bring great business value to organizations and individuals equally. Added to that, prediction has been identified as a key research area of the future.

On the other hand, predictive analytics is differentiated from prescriptive analytics which refers to the determination of a course of actions or decisions. In other words, the focus of prediction is on what will happen, whereas the focus of prescription is on how to make it happen [16]. For example, in a telecommunications operator content, predicting works to identify which customer will churn, while prescription works in ways to avoid it from happening via say simulation models.

BDA challenges

Researchers and practitioners alike face various types of challenges when using big data analytics for prediction, for instance, privacy and security of big data [17], platform scalability, integration, etc. However, for our purposes we only focus on those challenges that require epistemological reflection due to the bias incorporated in current practice, namely 'streetlight' research and data monetization.

'Streetlight' research

Big data is being passively created and continuously collected, and this has opened the door for plenty of research to be conducted. However, research should be formulated around important problems [18]. Yet, it has been noticed recently that big data research may have suffered from the so-called 'streetlight' effect. That is, the tendency of researchers to study phenomena for which there exist plethora of data, instead of studying relevant problems. To explain, most of the experiments and data-analytic research is relying on data from biggest data-driven companies e.g., Facebook, Twitter, Google, LinkedIn and Amazon. Great percentage of such studies is focusing on the data made available for researchers by those companies, for internal purposes. That is, such data may be either biased towards solving those companies' problems, and not necessarily the grand problems.

For instance, [19] showed that Twitter has become one of the favorite BDA research destinations. Such choice (of Twitter) by researchers is justified by its relatively high-level of accessibility and the relative openness of its API. Together, such two factors, have led to a substantial number of studies dealing with Twitter data. However, regardless of the case, the relative ease of data collection and analytics always entails the risk, and bias, of 'streetlight' research in BDA.

The "We Are Social Report", 2016 digital Yearbook, ranks Twitter as 9th in popularity as a social platform with 320M users, while other platforms have almost double or triple the number, such as FB (1.5B), WhatsApp (900M), etc. Twitter is certainly not the largest pool of users, and some of the accounts are used by bots, not humans [9]. Furthermore, many companies are using it as a way to boost sales, analyzing tweets "only" is indeed biased. Lastly, research observed that Twitter not only enables effective broadcasting of valid news, but also of rumors; as a matter of fact, false rumors would spread more quickly [20].

Since researchers can only analyze existing data, many are tempted not to formulate a clear research question or problem that enables to define what data is needed. In consequence, the range of insights we could or are able to generate remains unconsciously limited.

Data monetization

Data monetization is the ability of a company to generate money from its available datasets (partially or as a whole). In today's environment, companies have become aware of the meaning of the term "data is the new oil". Accordingly, each company is sitting on sheer amounts of data that needs to be utilized towards value creation. The way data monetization is implemented at companies could either be direct or indirect. Direct data monetization means selling (part of) the dataset of a company as such. Indirect monetization uses the dataset to create new products and services, such as Amazon is using its customer records to suggest other products or Alibaba via its targeted finance. Another form of indirect monetization takes place whereby a company is bartering its datasets.

Researchers can access data from data-driven companies e.g., Twitter, Facebook, Google, etc. via two mechanisms: API (aka 'garden hose') or a 'firehose'. API, or application programming interface, is a tool created for developers to interact with data producers. For instance, Twitter has created an open API allowing developers to source

Twitter data. The major advantage of the API is to promote external innovation, based on data. Offering data externally allows developers to create products, platforms, and interfaces without the need to expose the raw data. As a byproduct, Twitter has capitalized on this model by the acquisitions of ten different companies in 2012, built around their open API.

The 'firehose' is closely similar to the streaming API. The Twitter firehose guarantees delivery of 100% of the tweets that match search criteria by researchers. Data providers like GNIP and DataSift handle Twitter firehose. The firehose consists of an agreement between researchers and distributor of the firehose e.g., GNIP on tweets the researcher should receive. As the data providers receive tweets they are pushed directly to the end user.

The Twitter API is offered for free, but the Twitter firehose, which removes a lot of the usage restrictions imposed by Twitter, comes at a fee that not all researchers could afford. That fee represents what is known as "data monetization" for Twitter. Of course, researchers need to delimit their scope based on the data available. The key issue here is to be aware of the limitations of the datasets and the tools employed and to detail one's research approach accordingly.

Epistemological pitfalls in the BDA process

Investigating the epistemological challenges and pitfalls is crucial to the IS community which is becoming more and more multidisciplinary as well as multinational [21]. Numerous authors have discussed the potential of BDA for IS research, for example, Shimueli and Koppius [8] described six roles for predictive analytics: generating new theory, develop measurements, comparison of competing theories, improvement of existing models, relevance assessment, and assessment of the predictability of empirical phenomena. Three of these are particularly facing epistemological pitfalls, those are generating new theory, improve existing models, and assess predictability of empirical phenomena.

Recently [9] have provided guidelines for employing BDA in IS research. They conclude that "reflecting on the guidelines, we can observe that each phase of the research process requires a revised set of actions and abilities" (p. 11) and advocate a skill set change for IS researchers with stronger emphasis on developing skills for data preparation and the deployment of analytical tools and cross-instrumental evaluation criteria.

However, we seek to go beyond previous work by scrutinizing in more detail the BDA steps of data acquisition, preprocessing, analysis, and interpretation in order to identify the epistemological challenges associated with BDA. Concerned with the theoretical knowledge needed to appropriately apply BDA within the frame of IS research we seek addressing the following practical questions that call for an epistemological reflection:

- What kind of data [or datasets] about the world are available to a data scientist or researcher?
- How can these data [sets] be represented?
- What rules govern conclusions to be drawn from these datasets?
- How to interpret such conclusion?

Before conducting the actual steps of analytics, the primary stage is to define an objective, or identifying a problem to solve or an opportunity to grasp. That prerequisite step helps defining what needs to be accomplished. Quite often, the researcher might have many competing objectives and constraints that need to be properly adjusted and balanced. The appropriate identification of the objective or goal supports obtaining the right data, which has cascading impact on the entire BDA process as data is linked to analytics and analytics outcome is linked to interpretation. Therefore, defining the objective of analytics usually influences the result of the BDA process, especially when generating new theory, improving existing models, and assessing predictability of empirical phenomena. Added to that, the primary objective is normally linked to other related questions that need to be addressed, too. For example, the objective of a specific Telecom Operator is to "predict customer churn" [which might: generate new theory or improve existing known models]. Related questions are for example: who are the most profitable customers? Which of the profitable customers are influencers? How many complaints do we currently get from customer segment "profitable"? What are the products and services used by our top-profitable customers? Etc.

A possible consequence of neglecting this primary stage is to spend resources on producing the right answers to the wrong questions. Also, not having a clear research objective or problem, researchers will not be able to define what data is required to be collected and are tempted to undertake 'streetlight' research (see above). Of course, such defiance is expected to harm any kind of research design, but epistemological pitfalls in BDA are different. In traditional deductive research the existing body of theoretical knowledge guides the identification of relevant constructs, relations, and variables, and therefore influences the data collection from the outset. The various forms of inductive research (e.g. action research, ethnographic research, grounded theory) also rely on certain well explained and reflected approaches to small-size sampling, data collection, and data analysis that are to be applied and balanced from the outset according to the primary research objective (e.g. descriptive, exploratory, explanatory). For BDA such reflection of research design does not yet exist. Aiming for data-driven (not theory-driven) discoveries, the best practice being applied in deductive and inductive research so far does not work in this case. Hence, we need to reexamine every step of the analytics process in order to understand what kind of theoretical knowledge may help in avoiding the appearing epistemological pitfalls.

Acquisition

Big data analytics starts with acquiring the data through copying, streaming etc. (see also "BDA challenges"). Such acquisition requires good understanding of the domain (often business context) as well as the data. Datasets, from which we source data, should be described in terms of: required data to be defined; background about the data; list of data sources; for each data source the method of acquisition or extraction; and reporting the problems encountered in data acquisition or extraction.

One of the challenges associated with big data acquisition is: on one hand, there exist too much data while, on the other hand, all acquisition requires time, effort and resources. As pointed out above, the selection by the researcher might be attributable to: personal preference; technical abilities; 'streetlight' effect; and/or data monetization

impact. In practice, researchers seek technological solutions, i.e. tools to acquire and compress the data, and focus on available data. However, such solutions do not really address the epistemological problem: we know that sampling in data collection is crucial and requires a great deal of reflection pertaining to the impact of data acquisition decisions on the result of the research. Similarly, big data acquisition entails epistemological problems that require epistemological solutions, which cannot be achieved without sufficient theorization.

Preprocessing

Preprocessing activities include: check keys, referential integrity, and domain consistency; identify missing attributes and blank fields; replacing missing values; data harmonization e.g., check different values that have similar meanings such as customer, client; check spelling of values; check for outliers. In result, preprocessing provides a description of the dataset including: background (broad goals and plan for pre-processing); rationale for inclusion/exclusion of datasets; description of the pre-processing, including the actions that were necessary to address any data quality issues; detailed description of the resultant dataset, table by table and field by field; rationale for inclusion/exclusion of attributes; and the discoveries made during pre-processing and their potential implications for analytics.

Preprocessing mainly aims for big data cleansing and harmonization, while quite often overlooking the importance of 'traditional' data collection by the researcher. Big data self-confidence tends to drive preprocessing towards the assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. The core challenge is that most big data in focus are not the output of instruments that were designed to produce valid and reliable data amenable for rigorous knowledge discovery.

For example, a Google Flu prediction error in February 2013 resulted in doubling the proportion of doctor visits for influenza-like illness in the USA [22]. In this case the initial error was a marriage between big data and small data. Quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data. It is to be noted here that any empirical research must stand on a foundation of sound measurement, which not only include the data, but also its preprocessing.

Analytics

Despite the significance of predictive analytics, empirical analytics research is still rare in the IS literature. Extant IS literature relies almost exclusively on explanatory statistical modeling, where statistical inference is used to test and evaluate the explanatory power of underlying causal models, and predictive power is assumed to follow automatically from the explanatory model [1, 8]. Having that being said, the central step in BDA is analytics during which data mining, machine learning, statistics and other techniques, or models, are chosen and applied on the data. For the implementation of a technique (or model) a number of algorithms are available to be applied to any dataset. For example, we have conducted an experiment on a retail chain on which we have analyzed 1 year of purchase transactions for possible unnoticed relationships between products that ended up in shoppers' baskets. Discovering correlations between certain

items uncovered hidden patterns, which helped marketing team to promote low selling together with high-selling items. In such experiment, there was no hypothesis that a certain product, e.g. 1000, has been often bought with another product, e.g. 2000. The data were simply queried to discover what relationships existed that might have previously been unnoticed.

It is important on such step to describe: model assumptions, model description (e.g. rule-based models list the rules produced in addition to their accuracy and coverage), and results assessment (e.g. why a certain modeling technique and certain parameter settings led to good or bad results).

One efficient approach to follow in BDA research is to identify, early enough, what one is looking for. However, data scientists responsible for the analytics process are often not aware of this epistemological challenge (i.e. know what you want to know) and/or do not have the necessary knowledge to apply in that manner. Hence, it turns out that preferences of the data scientists and their education might drive the analytics part instead of the problem at hand, leading to insufficient knowledge discovery.

In big data analytics, the variables represent the raw input whereas the feature is a variable selected or (re)constructed from raw variables. Hence, feature selection is part of pre-processing. It helps to reduce the measurement and storage requirements, reducing training and utilization times. It addresses the high-dimensionality problem. The idea is to selecting the best features that are useful to build a good predictor. This is not the same problem as finding or ranking all potentially relevant variables. Selecting the most relevant variables is usually suboptimal for building a predictor. The relationship between feature selection and model predictive accuracy should be emphasized.

In particular, the selection of algorithm's parameters by the data scientist has a profound impact on analytics. A parameter is a value to fine-tune an algorithm. For any BDA tool e.g., RapidMiner, there are often a large number of parameters that can be adjusted. Listing the parameters and their chosen values, along with the rationale for the choice, is a key task. For instance, for the K-Means clustering algorithm, setting the number of k is a parameter. Too big k might not be useful for the decision maker and too little value for k as a parameter might not solve business problems. While empirical research stands on a solid foundation of measurement, data scientists tend to overlook the fact that algorithms parameter setting not only impacts analytics, but interpretation as well.

The algorithms that preprocess and analyze big data to find patterns, trends, and relationships are in many cases treated as 'black-boxes' or 'closed'. Yet, understanding analytics algorithms is of importance because they not only extract and derive meaning from the world, but they are increasingly starting to shape it. However, in many cases, that shaping is semantically blind. For instance, Google matches ads to content without 'knowing' anything about either. The Google translate service (and the team behind it) does not understand content of the language they are providing translating for. Netflix reported that 75% of content choices made by their customers is influenced by their recommendation system [23].

The following two techniques may illustrate the difficulties of managers and users to understand analytics:

Support vector machines (SVM)

Discovering the right set of features is a difficult problem in machine learning. SVMs try to model such feature list. The idea of SVM is to make use of a [nonlinear] mapping function Φ, which transforms data in input space to data in feature space in such a way so as to render a problem linearly separable whereby the SVM is then able to automatically discover the decision surface (DS). There are plenty of ways where DS's could be identified, see three potential lines on Fig. 1.

SVM then automatically discovers the optimal separating hyperplane which, when mapped back into input space via $\Phi - 1$, could be a complex DS. The discriminating hyperplane in input space corresponds to the function:

$$\varpi = \sum_i \alpha_i \tilde{\mathcal{S}}_i$$

where the ϖ is the omega vector; and the RHS represent the Sigma of Lagrange parameters [AKA alpha α_i parameters] over S vectors treated for input bias—the $\tilde{\mathcal{S}}_i$.

Ensemble methods: random forests

Ensemble methods use a divide-and-conquer tactic used to improve performance. The core principle is that a cluster of weak learners[1] could come together to procedure a strong learner.

The idea of ensemble methods is illustrated in Fig. 2 whereby each classifier individually is a weak learner, however, when taken together the classifiers represent a strong learner. The data to be modeled are represented by the blue circles. Each learner model is represented as a gray curve. Each gray curve is a fair approximation to the underlying data. The red curve represents the assembled strong learner model; which could be seen as a better approximation to the data. Random forests is based on tree induction (aka decision trees) and is frequently used in prediction, where one needs to know: bagging, pruning, cross-validation, entropy measures e.g., Gini index, etc. in order to fully understand how it works and being able to digest its results.

Improving prediction accuracy could be achieved by using ensembles. Ensembles means averaging across multiple models that rely on different data or reweighted data and/or employ different models or methods. Bagging, boosting, and random forests are frequently used ensemble methods. Ensemble method require voting. The voting operate on class labels, where $d_{t,j}$ is 1 or 0 depending on whether classifier t chooses j, or not, respectively. The ensemble then chooses class J that receives the largest total vote. In ensemble methods, in order to combine the classifiers, boosting takes a weighted majority vote of their predictions. On the other hand, bagging uses bootstrap samples to build the classifiers. Each bootstrap sample is constructed by randomly sampling, with replacement, the same number of instances as the original data. The final classification produced by the ensemble of these classifiers is obtained by simple majority voting.

[1] In machine learning, an algorithm that learns from data is called learner models. Such algorithms build a model from training data in order to facilitate classification and predictions. A core objective of a learner is to make generalizations from the data. Generalization is the ability of a learner to predict accurately on unseen examples. Since training data came from unknown probability distribution, the learner has to build a general model that enables it to produce accurate predictions in new cases.

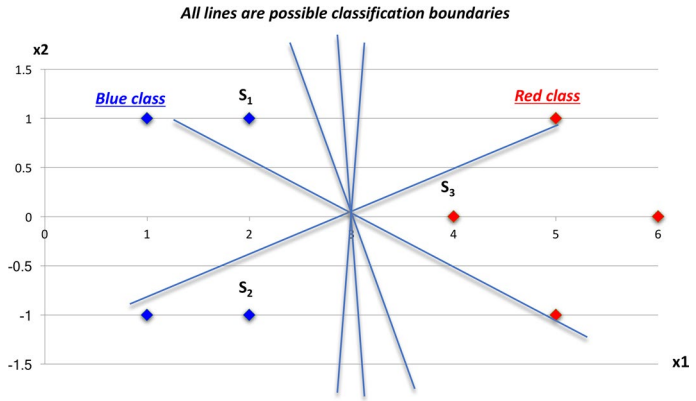

Fig. 1 SVM options for decision surface discovery

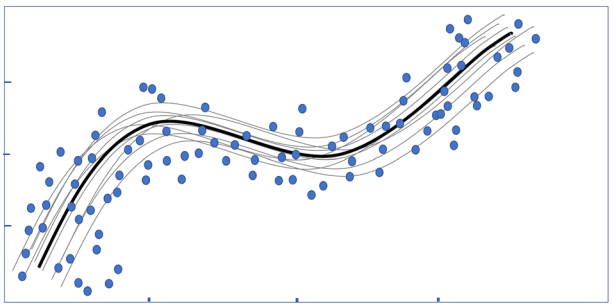

Fig. 2 Ensemble methods: random forest

Transparency in data collection, preprocessing, and analytics (esp. parameter setting) is inevitable in data science. The above techniques are just brief examples for the powerfulness but also for the complexity of BDA. If data scientists would not understand and know how they generate predictions, then we are unable to address epistemological issues in the BDA process. The human element of big data (analytics) is strategically important. That is, it is indeed essential to combine potentials of machine learning algorithms with human decision making skills. There is still a gap between what machine learning and statistics tools used in big data analytics could do, and what could be done with this generated knowledge i.e., role of human [24].

Interpretation

Interpretation should relate analytical findings to the existing body of knowledge as well as industry practices and include reflection on certain business objectives, decision making, problem solving, etc.

One of the significant epistemological problems in this step is the interpretation of 'quick & dirty' pattern discovery. The reason is mainly attributable to the fact that analytics can run easily and quickly by the data scientist, even via cloud. Given these opportunities, the pressure to reach outcomes often supersedes the genuine objective of the advancement of knowledge.

Another issue is the contradiction of predictive and explanatory power. Often, BDA provides us with higher accuracy prediction, but this accuracy comes at a cost. That is, most accurate algorithms such as SVM, Naïve Bayes Classifiers, topic modeling in text analytics, and random forests are not easily comprehensible by most of those who are supposed to consume their results i.e., decision makers and managers (see also "Analytics"). In other words, BDA utilizes algorithms that are good in predicting future or unknown events, but unable to provide easy-to-comprehend explanations for their predictions. On the other hand, many decision makers and managers have learned to interpret regression results. Therefore, we are facing a situation whereby, when users are to choose between accuracy (of BDA algorithms) and interpretability (of 'traditional' statistics), many would favor interpretability [9].

One example here is the research conducted by [25] which has resulted in the construction of a corpus of digitized books containing 4% of the books ever printed, a corpus of 5.2 million digitized books. Applying analytics on such big data enables better understanding of cultural trends that have prevailed in history. Researchers have used such big data corpus in order to understand grammar, collective memory, technology adoption etc. Such corpus is a result of Google efforts to digitize books. Having that being said, one should be very conservative in interpreting results obtained from such corpus since the corpus contains >500 billion words but not equally represented languages. That is, it has such number of words per language: English 361B, French 45B, Spanish 45B, German 37B, Chinese 13B, Russian 35B, and Hebrew 2B. Added, the corpus was collected from approximately 40 university libraries worldwide i.e., not a large representation.

Big data analytics uses various techniques, such as machine learning techniques, to identify the likelihood of future outcomes based on applying those techniques on available datasets. Those datasets are generated from a variety of sources having different representation forms and formats. Added to that, the techniques have assumptions and parameters. All of that raises risks of (mis)interpretation and hence render the business decision made based on BDA findings invalid! Therefore, businesses utilizing BDA outcomes should investigate further the steps of BDA in order to safeguard their knowledge discovery activities as well as their fact-based decision making. Addressing this gap, we introduce a theory-driven guidance to avoid the epistemological pitfalls and to help mitigating the epistemological challenges encountered during the BDA process.

Addressing the epistemological challenges

Trying not to let BDA fall into empiricism, we looked at each BDA step and the related critical question in performing this step in order to identify the main epistemological challenges. In result we recommend a "lightweight theory-driven" approach in contrast to a "heavyweight theory-driven" research that is solely based on popular or relevant theories pertaining to the research. The advantage of the latter is the ability to derive generalizable research outcomes that are easily interpreted and compared. The disadvantage is that conflicting theories could exist to choose from which makes it unclear whether a selected theory would hold in the application domain.

For example, in consumer perceived value research, a heavyweight research would have followed known theoretical frameworks such as [26] identified four dimensions:

quality, emotional, price and social. However, pursuing a BDA process on 18,000 Amazon customer reviews on two product categories (cameras and tablets) revealed, based on topic modeling analytics, four dimensions for the cameras (emotions; features; post-sales services; events) and five dimensions for the tablets (feedback, gaming, post-sales services, features, and battery/charging). This shows that the BDA process has identified new dimension e.g., post-sales services, and informed us that dimensions could vary by the product categories. Such finding helps retailers and other stakeholders to predict customer reaction, based on the newly identified value dimensions.

Avoiding a theoretical commitment from the outset, a lightweight theory-driven BDA may start by data acquisition followed by the remaining steps. In each, there exist some recommendations in order not to wait for competing or conflicting theories. Lightweight theory-driven acquisition and preprocessing consists of activities such as data summarization, graphical representation, dimension reduction; and outlier detection. Dimension reduction means reducing the number of dimensions is normally accomplished via methods such as principal components analysis (PCA).

Lightweight theory-driven data and parameter selection for knowledge induction means relying on body-of-knowledge and existing theories in order to go beyond a mere quantitative analytics approach. For instance, one way to do this is to map the constructs of analytics with known theoretical constructs. Preferably, multiple researchers contribute to constructs' identification and cross-disciplinary contribution in the mapping process. The rationale behind being multi-disciplinary is that an a priori knowledge of the data and the represented domain helps in sharpening multi-dimensional understanding of the analytics as a process and hence pave the road for sound conclusions that contribute to science and practice. However, this can be considered 'lightweight' because reaching out to the data does not start hierarchical from the theoretical concepts and constructs and respective variable definition, but from the given datasets mainly 'to sort things out'.

Examining data quality, validity and reliability, for example in data warehousing, usually includes questions such as: Is the data complete? Is it correct? Does it contain errors, and if there are errors how common are they? Are there missing values in the data, and how common are they? In BDA, similar efforts are required, but to scale. In particular, the assessment of big data quality should be made not just of the individual data source (e.g. OLTP systems), but also of any data that came from merging sources. This is due to the fact that merging data is quite often the case in big data projects, which causes potential problems, e.g. inconsistencies between the sources, which do not exist in individual data sources. Also, in data warehousing the data is modeled upfront, i.e. before utilization, which makes it necessary to build extract-transform-load (ETL) before use and to apply required quality and validity checks. However, in BDA the data is modeled afterwards, and that makes it different in terms of when data validity and reliability takes place. In BDA research, attention to validity and reliability is required because the scale of the datasets is often huge, the variety of data types is high, and the model is built after the data has already been collected—and because overlooking data validity and reliability issues would risk ending up with contaminated analytics and hence interpretations.

Finally, a theoretical framework should govern BDA because complacency about the modeling technique causes epistemological problems in result interpretation. Without

such framework the selection of the technique might be mostly based on: tool availability, knowledge of the data scientist/researcher, and/or being politically friendly to stakeholder expectations, with all the problems that might arise from such bias. Instead, models should be selected based on: given data, problem at hand, and model assumptions. For instance, we should not use correlation in a problem for which we need to know cause-and-impact, since correlation coefficient does not imply causality. It goes for association rules, which indicate frequency and occurrences, but have least prediction power. Also, we should use SVM when we have nonlinearity and selection of features is required.

Table 1 summarizes this effort and indicates possible theoretical contributions (explained below) that could guide through mastering the identified epistemological challenges.

Data summarization

The idea of data summarization is simple. For example, in order to understand the relationship between qualifications and income, dataset could be viewed by plotting the average income by qualifications level. Such summary will be sufficient for some purposes, but if the outcome of summarization is used in fact-based decision making, more time is required in order to achieve a better understanding of the data. A simple example would be to include the standard deviation information along with the averages. Further, it may be more revealing, for example, to break down the average income levels by age group, or to exclude outlier incomes. Moreover, the relationship between income and qualifications may vary between men and woman, or may vary by geography. Overall, effective summarization involves both identifying overall trends and important exceptions to them.

Table 1 Lightweight theory-driven guidance for the BDA process

BDA step	Critical questions	Epistemological challenge	Possible lightweight theory-driven guidance
Acquisition	What data do I need? What kinds of data [sets] are available/to be selected?	Data 'sampling'	Apply data summarization, graphical representation, dimension reduction (e.g. PCA) and outlier detection
Pre-processing	How can data [sets] be represented and processed without falsification or insight loss?	Data validity and reliability	Ensure multi-expert and multi-disciplinarily participation in data reduction and selection Trace and examine all stages of extract, transform, load, and merge for completeness, correctness, and consistency
Analytics	Which method[s] to use? What rules govern conclusions from these data [sets]?	Knowledge discovery	Map the constructs of analytics to known theoretical concepts Ensure multi-expert and multi-disciplinarily participation in parameter selection and mapping analytical constructs with theoretical concepts Develop/apply theoretical framework for choice of techniques (mining, machine learning, statistics) or models
Interpretation	How to interpret such conclusion?	Non-/interpretability; reliability of prediction	Develop/apply theoretical framework for result interpretation

Graphical representation

Graphical techniques aid users in managing and displaying data in an intuitive manner. Visualization can be helpful in the discovery of relationships and dependencies that may exist within the dataset. The core issue here is to effectively representing multidimensional datasets without overwhelming the human ability to comprehend the resulting graphs. Data summarization can reduce the size and complexity of multidimensional datasets. This could highlight the relevant aspects of the dataset more clearly, leading to more coherent visualizations, and also facilitating more accurate and efficient visual analytics.

Outliers detection

In the following figure, if we look at the top-right most point, it seems like an outlier, since we are looking with regards to how far from rest of the data points which are depicted in two axes. However, if we look at such point with regards to x-axis or y-axis only, it will not be identified as outlier. Outlier detection techniques can be categorized, based on the number of variables or dimensions used to define the outliers, into two categories; univariate outlier detection techniques; where the outliers are detected for only one variable at a time, and the other category is the multivariate outlier detection where more than one variable is taken into account while defining the outliers. Most probably the univariate outlier detection is insufficient. So, in BDA analyzing the dataset based on univariate outlier detection method leads to epistemological pitfall. Correction is to rely on multivariate techniques in outlier detection [27] (Fig. 3).

Dimension reduction

Dimension reduction refers to the process of converting a dataset of high dimensions, into dataset with less dimensions. However, similar information should be precisely conveyed. Dimension reduction techniques are used in BDA in order to obtain better features for a classification or regression task. We can reduce n dimensions of dataset into k dimensions (where $k < n$). The k dimensions can be directly identified or can be a combination of dimensions (weighted averages of dimensions) or new dimension(s) that represent existing multiple dimensions. Dimensionality reduction takes care of multicollinearity, by removes redundant features. That is, variables exhibiting higher correlation, which could cause low predictive power of a model. Both factor analysis and principal component analysis are used for the purpose of dimension reduction. Factor Analysis, assumingly the dataset used in analytics has highly correlated variables. These variables can be grouped by their correlations whereby variables in a group are highly correlated, but exhibit low correlation with variables of other groups. Here each group represents a single underlying construct or factor. These factors are small in number as compared to large number of dimensions. On the other hand, using principal component analysis (PCA), variables are transformed into a new set of variables, which are linear combination of original variables. These new set of variables are known as principle components. They are obtained in such a way that first principle component accounts for most of the possible variation of original data after which each succeeding component has the highest possible variance. The second principal component must be orthogonal to the first principal component. In other words, it does its best to capture the variance in the data that is not captured by the first principal component. For two-dimensional dataset, there can be only two principal components.

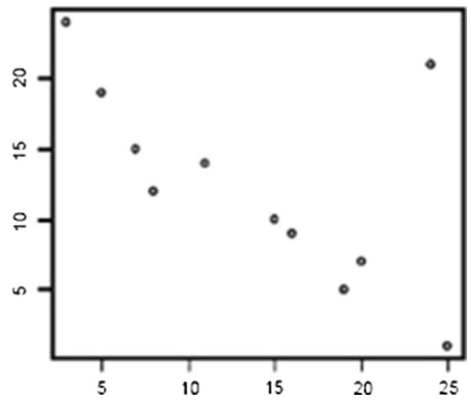

Fig. 3 Outlier detection

ETL

The ETL refers to the extract, transform, and load process. Normally, this is associated with the data coming to the data warehouse form multiple source systems [28]. During extract, variables are extracted from various sources e.g., an ERP database. In the transform, data is transformed into a desired structure e.g., source systems have price and quantity while target systems—the data warehouse—have total i.e., multiplication of price and quantity. The transformation may also include filtering unwanted data, sorting, aggregating, joining data, data cleaning, data validation based on the business need. Lastly, the load step involves the transformed data being loaded into a destination target, which might be a database or a data warehouse. One of the issues that might occur during ETL is improper, or incomplete mapping. That is, not all columns in source systems are mapped to destination systems. If happened during data acquisition step of the BDA, that will lead to data loss or errors hence impact knowledge outcome and model predictive power.

Merge for completeness

Big data is generated at different source systems, so bringing such data together is a challenge. For example, in order to ensure completeness of the dataset in BDA, the data is to be retrieved from n sources: S1, S2, …, Sn. All source systems must send their data records to the central repository and it should be possible to define a relational operation that will reconstruct the dataset from the multiple sources. Such reconstruction could be horizontal via Union operator and vertical via Natural Join operator. For example, in case the dataset is split horizontally, we could retrieve them all in the central repository—data warehouse—using selection operator. That is, if the dataset comes from two horizontal sources, we then need two statements for full reconstruction: σ type = 'A' (S1) $\cup$ σ type = 'A' (S2). Or vertical sources reconstructed: Π ProductNo (S1) $\cup$ Π ProductName (S2). Failing to reconstruct the dataset completely leads to prediction accuracy being low for the model used [29].

Fundamentally, BDA automates the knowledge discovery process from data, or datasets, in order to make predictions. Such discovery is a genuine machine science in which all process steps are subject to automation. Generating new theories is among the roles

predictive analytics is expected to play (see above), achieved through the development or improvement of models. Such theories target to predict a variable in the future, given a set of explanatory variables or predictors. Some theories may even be able to explain the causal relationship between independent and the dependent, while others do not have such explanatory power.

One of the core question for science and practice regarding utility is: *what are the necessary epistemological preconditions that make predictions based on model-based data processing acceptable for human stakeholders?* Here we consider the primary criterion to be the performance: the success of the prediction (i.e. it turns out to be true) is far more important than how we have reached it. Our rationale here is grounded on the counter question: *What else could be more relevant to assess prediction rather than its correctness?* If theory is not able to correctly predict the future, we also start to question every aspect of it (constructs, variables, relationships, assumptions, context) and/or leave it aside.

There is an argument that a prediction without explanation is inferior, hence BDA-based predictions lack the explanation power in many occasions and that render them 'incomplete'. However, often enough predictions based on explanatory theories are not accurate (enough). And in many real life cases prediction precedes explanation, i.e. certain phenomena (e.g. an epidemic or seasonality sales) can and/or should be predicted with the reason behind the phenomena to be revealed only later (if at all). It is certainly the case that to some extent BDA-based prediction lacks the explanatory power. But especially in the beginning of exploring new phenomena this should not be considered as an argument to rule out BDA application. And in such cases the acceptability of the prediction and the trust in its results should rather be derived from the transparency and lightweight theory-driven governance of the BDA process.

Conclusion

Indeed, BDA seeks to gain insights "born from the data" and entails "disruptive innovations" [2] with implications how research is conducted: instead of inductively proposing theories from small sample data and/or deductively confirming theories based on theory-driven instruments and data collection, we now process given big datasets step by step to generate relational insights and predictions. The core of the shift pertains to the scientific method employed in BDA. In BDA the research can start with processing huge amount of data to reach data-driven discoveries, rather than starting with theory or with small sample data to be interpreted by humans.

Nevertheless, scientific theories have not become obsolete in BDA research. But the shift towards process-driven generation of insights and predictions poses new epistemological challenges that require different theoretical guidance for each step in the data processing, particularly lightweight theory-driven data and parameter selection, systematic big data validity and reliability reflection, and an overall theoretical framework supporting method selection and result interpretation. All of which should be on the agenda of IS research and need to be addressed by existing and/or to be developed IS theories.

Recently, few research studies, (namely [2, 9]) have introduced guidelines on how to conduct big data analytics research. However, triggered by the propagating challenges of 'streetlight' research and data monetization, our study differs from the previous papers

by focusing on epistemological challenges: having identified the possible epistemological pitfalls in each step of the BDA process, we have introduced a lightweight theory-driven guidance that aims to improve the governance, acceptability, and eventually trustworthiness of BDA.

This work is only another stepping stone towards reflecting and addressing the epistemological challenges associated with BDA and BDA-based predictions. Further research is required with regards to the impact and support of BDA-based predictions in relation to IS theories as well as management theories. In particular, longitudinal studies on the relation of explanations and predictions are needed in order to appropriately contextualize BDA in the history of science.

Authors' contributions
Both authors worked on the preparation of the manuscript. AE focused more on the big data analytics part and the challenges of the prediction process. RK focused more on the theoretical background and epistemological pitfalls and how to address them. Both authors read and approved the final manuscript.

Author details
[1] Department of Computer Science, Electrical and Space Engineering, Computer and Systems Science, Luleå University of Technology, 971 87 Luleå, Sweden. [2] Faculty of Management Technology, German University in Cairo, Main Entrance Road-Fifth Settlement, New Cairo City 11835, Egypt.

Acknowledgements
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 22 January 2017 Accepted: 8 June 2017
Published online: 23 June 2017

References

1. Dhar V. Data science and prediction. Commun ACM. 2013;56(12):64–73.
2. Kitchin R. Big data, new epistemologies and paradigm shifts. Big Data Soc. 2014;1:1–12.
3. Feyerabend P. Against method: outline of an Anarchist theory of knowledge. London/New York: New Left Books; 1975.
4. Bacharach S. Organizational theories: some criteria for evaluation. Acad Manag Rev. 1989;14(4):496–515.
5. Whetten D. What constitutes a theoretical contribution. Acad Manag Rev. 1989;14(4):490–5.
6. Colquitt J, Zapata-Phelan C. Trends in theory building and theory testing: a five-decade study of the Academy of Management Journal. Acad Manag Rev. 2007;50(6):1281–303.
7. Gregor S. The nature of theory in information systems. MIS Q. 2006;30(3):611–42.
8. Shimueli G, Koppius O. Predictive analytics in information systems research. MIS Q. 2011;35(3):553–72.
9. Müller O, Junglas IA, Brocke JV, Debortoli S. Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur J Inf Syst (EJIS). 2016: 1–14.
10. Agarwal R, Dhar V. Big data, data science, and analytics: the opportunity and challenge for IS research. Inf Syst Res. 2014;25(3):443–8.
11. Chen G, Guo X. Big data commerce. Inf Manag. 2016.
12. Elgendy N, Elragal A. Big data analytics: a literature review paper. The 14th Industrial Conference on data mining (ICDM). Petersburg: Springer-LNCS; 2014.
13. Russom P. Big data analytics. TDWI 4th Quart. 2011:1–38.
14. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev. 2012;90:3–9.
15. Fayyad U, Piatetsky G, Smyth P. From data mining to knowledge discovery: an overview. In: Fayyad U, Piatetsky G, Smyth P, editors. Advances in knowledge discovery and data mining. Cambridge: AAAI Press/The MIT Press; 1996. p. 1–34.
16. Provost F, Fawcett T. Data science for business. Newton: O'Reilly; 2013.
17. Newell S, Marabelli M. Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of 'datification'. J Strateg Inf Syst. 2015;24:3–14.
18. Rai A. Synergies between big data and theory. MIS Q. 2016;40(2):iii–ix.
19. Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D. The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int J Commun. 2011;5:31.

20. Mendoza M, Poblete B, Castillo C. Twitter under crisis: can we trust what we RT? In Proceedings of the First Workshop on social media analytics (SOMA), Washington DC, July 25–28, 2010, p. 71–79.
21. Becker J, Niehaves B. Epistemological perspectives on IS research: a framework for analysing and systematizing epistemological assumptions. Inf Syst J. 2007;17:197–214.
22. Lazer D, Kennedy R, King G, Vespignani A. The parable of google flu: traps in big data analysis. Science. 2014;343:1203–5.
23. Lycett M. 'Datafication': making sense of (big) data in a complex world. Eur J Inf Syst. 2013;22:381–6.
24. Depeige A. Taming the realm of big data analytics: acclamation or disaffection? In: Tomar GS, Chaudhari NS, Bhadoria RS, Deka GC, editors. The human element of big data: issues, analytics, and performance. Boca Raton: CRC Press; 2016.
25. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, et al. Quantitative analysis of culture using millions of digitized books. Sci Express. 2010: 1–11.
26. Sweeney J, Soutar G. Consumer perceived value: the development of a multiple item scale. J Retail. 2001;77(2):203–20.
27. Badawy S, Elragal A, Gabr M. MSCM: an outlier detection technique for data mining applications, In proceedings of the conference on artificial intelligence and applications (IASTED-AIA), Innsbruck, 11–13 February 2008, p. 314–20.
28. Sharda R, Delen D, Turban E. Business intelligence, analytics, and data science: a managerial perspective. 4th ed. Upper Saddle River: Pearson; 2017.
29. Connolly T, Begg C. Database systems: a practical approach to design, implementation, and management. Upper Saddle River: Pearson; 2014.