Decision Support Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A snail shell process model for knowledge discovery via data analytics

Yan Li [a,*], Manoj A. Thomas [b], Kweku-Muata Osei-Bryson [b]

[a] Claremont Graduate University, Center for Information Systems and Technology, 130 E. 9th Street, ABC 217, Claremont, CA 91711, USA
[b] Virginia Commonwealth University, School of Business, PO Box 84000, Richmond, VA 23284-4000, USA

ARTICLE INFO

Article history:
Received 9 September 2015
Received in revised form 8 July 2016
Accepted 15 July 2016
Available online xxxx

Keywords:
Knowledge discovery via data analytics
Snail shell process model
KDDA
Big data analytics
Data-driven decision making

ABSTRACT

The rapid growth of big data environment imposes new challenges that traditional knowledge discovery and data mining process (KDDM) models are not adequately suited to address. We propose a snail shell process model for knowledge discovery via data analytics (KDDA) to address these challenges. We evaluate the utility of the KDDA process model using real-world analytic case studies at a global multi-media company. By comparing against traditional KDDM models, we demonstrate the need and relevance of the snail shell model, particularly in addressing faster turnaround and frequent model updates that characterize knowledge discovery in the big data environment.

1. Introduction

Integration of technologies such as cloud computing, social networking, and mobile technology into business functions has propelled the creation of large volumes of data at high velocity from a variety of sources. Commonly known as big data, modern organizations recognize it as a valuable asset, and are increasingly attracted to the possibility of creating competitive advantage through data driven knowledge discovery. Yet, they grapple with the challenges imposed by the very nature of the big data environment such as, managing data, extracting information, and discovering knowledge. The large volume poses technological challenges in applying advanced analytics on big data platforms compared to the use of traditional SQL analytics directly on databases [11]. The high velocity calls for much faster and more frequent turnaround for knowledge creation. The large variety of data sources introduces enormous data integration and governance responsibilities. In an effort to address these big data challenges, industry trend now leans towards utilizing information technology (IT) and advanced analytic techniques for faster, cheaper, more flexible, and more reliable knowledge discovery.

For successful creation of useful knowledge, a comprehensive framework that describes how to carry out the analytic process is essential. The current approach for practitioners is to adopt traditional knowledge discovery and data mining (KDDM) process models for the integration of very technical analytic solutions into organizational business processes [28]. While significant progress has been made in big data analytics tools, methods and algorithms (e.g., Microsoft Cortana Analytics Suite, IBM Watson Analytics, and Teradata Aster), generalizable process knowledge for conducting knowledge discovery via data analytics (KDDA) in an organizational context has yet to be updated. For example, Netflix shared their analytics process to ensure consistencies in building predictive models [11]. However, it covers only a small portion of the complete analytic project life cycle. Furthermore, a review of literature indicates that existing KDDM process models [10,18,28,31,40] were developed prior to the popularity of big data and therefore not well suited for addressing many challenges unique to big data analytics. Their limitations include the lack of scalability and agility in the development of analytic solutions, longer cycles between data acquisition and decision making, insufficient considerations of organization's analytics capability maturity (ACM), and missing model maintenance components.

The purpose of this study is to investigate challenges of undertaking analytic projects in big data environment. It aims to improve deficiencies in existing KDDM models to provide decision support for different decision makers, such as business domain experts, data engineers, and analytics practitioners, throughout the entire KDDA process. Many studies have discussed of techniques currently employed in big data analytics [7,20]. Literature also overviews emerging and future trends in big data analytics [5,24]. Newer studies have investigated the impact of analytics on business function performance [38,46] and capabilities of analytic techniques [13]. Yet, no study has holistically addressed KDDA as an end-to-end process in a business context. To bridge this gap, this study seeks to draw upon existing literature to develop and formalize the systematic process of KDDA.

Our study makes three important contributions. First, by framing distinctions between traditional data mining projects and KDDA in a big-data-driven decision making environment, we identify crucial

* Corresponding author.
E-mail addresses: Yan.Li@cgu.edu (Y. Li), mthomas@vcu.edu (M.A. Thomas), kmosei@vcu.edu (K.-M. Osei-Bryson).

http://dx.doi.org/10.1016/j.dss.2016.07.003
0167-9236/

steps missing in commonly used KDDM models, and outline a much needed revision. Our study thus contributes to the wider body of scientific knowledge that has so far not holistically explored KDDA in a business context. Second, drawing upon published literature, we propose a KDDA process model, which we call the snail shell model. The model integrates eight key phases and related tasks at the meta-level, and draws attention to the highly iterative, yet interlinked phases. Finally, by using real-world analytic cases at a global multi-billion dollar technologically focused mass media company, we evaluate the effectiveness and relevance of the proposed model in supporting organizational knowledge discovery. The proposed model thus contributes to practice by supporting practitioners to carry out the KDDA process more effectively. Our proposed model not only prepare practitioners in making informed decisions in performing analytics tasks, but also allude them to the importance of assessing other related aspects, such as organization's ACM, data quality and governance, institutionalization of model management in the big data environment.

The rest of the paper is organized following Gregor and Hevner's [19] design science research study publication schema. Section 2 presents a review of the relevant literature and offers a comparative analysis of existing KDDM models with justifications for a KDDA process model. Section 3 describes the research method and study background. Section 4 provides a detailed description of the artifact, the snail shell KDDA process model, and compares it with traditional KDDM models. Section 5 presents the evaluation of the snail shell model which involved a real world big data analytics project. Section 6 discusses research implications and limitations, followed by conclusions in Section 7.

2. Need for a KDDA process model

Data analytics can be defined as "the analysis of data, using sophisticated quantitative methods, to produce insights that traditional approaches to business intelligence (BI) are unlikely to discover" [39]. In addition to the traditional data mining methods (e.g., tree induction, cluster analysis, and association rules), data analytics also include a wide range of quantitative methods (e.g., simulation and optimization), and visualization techniques. The integration of advanced data mining techniques and data analytics has led to the emergence of a new sub-discipline within IS called data science [14]. Provost et al. [37] define data science as "a set of fundamental principles that support and guide extraction of information and knowledge from data". In this paper, we use the term KDDA to describe the knowledge discovery process and practices in the analytic environment of an organization.

Knowledge discovery is a non-trivial process that requires not only technical knowledge of IT, analytic techniques, and mathematical algorithms, but also a thorough understanding of the business process. Currently, practitioners tend to adopt traditional KDDM models to organize analytics projects and communicate solutions to business users. A comparative analysis of fourteen popular KDDM process models by Mariscal et al. [31] revealed the 9-step Knowledge Discovery in Databases (KDD) process [18] as an initial approach and the CRISP-DM model [40] as a central approach. These KDDM process models share several commonalities (e.g., similar sequence of phases and steps, iteration of phases and feedback loops). Phases common among them are: business understanding (BU), data understanding (DU), data preparation (DP), data mining (Modeling), evaluation, and deployment. However, existing KDDM models do not reflect many changes that have occurred to KDDM applications in recent years, especially with the proliferation of big data in business processes. For example, 9-step KDD process [18] focuses on data transformation, but does not address capturing knowledge from KDD for reuse. Several other models [3,4] are very similar to the 9-step KDD approach and have similar limitations.

As a de facto industry standard, CRISP-DM focuses on ensuring quality of data mining projects by describing activities that must be done during the KDDM process. Its objectives are to reuse the process knowledge and reduce skills required for KDDM. However, CRISP-DM assumes a waterfall life cycle for data mining projects. It does not address key project management activities, such as quality management or change management [30]. The use of CRISP-DM has recently witnessed a decrease due to competing in-house methodologies developed by KDDM project teams, such as SEMMA (sample, explore, modify, model assess) by SAS institute [31]. However, in-house methodologies lack generalizability (i.e. stability across varying applications and insensitive to changes in the environment), and tend to be tools and techniques dependent. Mariscal et al. [31] proposed a Refined Data Mining Process that included three high-level processes (analysis, development, and maintenance) and 17 sub-processes based on the synthesis of existing approaches. Nevertheless, they focused on the description of sub-processes without providing a concrete description of "how to do" (i.e., tasks carried out in each sub-process).

The rapid growth of big data environment imposes new challenges that traditional KDDM models are not adequately suited to address. The pressing need for updating the existing KDDM process models are attributable to the following key issues pertinent to big data environments: 1) the large volume, high velocity, and wide variety of data demand scalability of analytic solutions and deployment [23]; 2) the increasing scale of KDDA projects results in the increasing reliance on teams, making it important to educate greater numbers of people on relevant analytic processes and best practices [8]; 3) the need to shorten the time between data acquisition and decision making [6] posits packaging analytic tasks for non-analytic end users and integrating these tasks in business workflows; 4) analytic models are knowledge intensive products that are not only expensive to build, but also expensive to maintain and deploy rapidly [27]; 5) the existing process follows the SDLC methodology where most business requirements are gathered at the beginning of the project, whereas most KDDA projects usually start with ill-structured business problems that requires more frequent and multiple iterations [45]; and 6) the organization's analytics capability maturity directly influences its success with actionable analytic solutions [22]. Based on existing literature, we propose an updated snail shell KDDA process model that caters to knowledge discovery needs in big-data-driven decision making environments.

3. Research method and study background

This research undertakes the design science paradigm as guided by the information systems research framework proposed by Hevner et al. [21]. The framework emphasizes the needs to achieve IS research relevance by framing research activities to address business needs in the appropriate environment, and research rigor by appropriately applying existing foundations and methodologies from the knowledge bases. Specifically, our research activities were carried out as an iterative build-and-evaluate process through exploration of a real-world analytic environment and knowledge bases of KDDM literature.

According to Hevner et al. [21], environment is the problem space where the phenomena of interest resides, and is composed of the organization, the existing (or planned) technologies, and its people. The organization where we conducted this research is a new product division of global multi-billion dollar mass media company. The division had recently released a new product, SmartBoxOne (SBO), which was viewed as the company's strategic move towards its high-tech service offering. From a technology perspective, the product division was an early adopter of big data platforms, and utilized many big data technologies (e.g., Splunk, Cassandra, Apache Flume, and Pig) to store and analyze data. In 2014, when this research was initiated, the division was indexing 7–8 TB of SBO log data per day. From the people's perspective, the division had a BI team that was responsible for SBO reporting for multiple business units. The BI team included a team of data engineers who were responsible for data extraction, transformation, and loading (ETL), a team of analysts who focused on reporting, and an executive

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

director. The director had realized that the team would soon exceed its capability to support the exponential growth of SBO. Hence, three researchers with established background in data management and analytics were invited by the director to assist in improving the analytic process. The researchers joined the team in mid-2014, and were actively involved for 14 months.

Hevner et al. [21] describe knowledge base as foundational theories, frameworks, instruments, constructs, methodologies and instantiations that provide raw materials from and through which IS research is conducted. To identify various phases of the analytic process, researchers relied on published frameworks and methodologies (e.g., CRISP-DM, SEMMA). To develop and refine analytic tasks and guidelines, researchers utilized relevant literature and their real-world experience in analytics. In the first six months, researchers held weekly meetings with the BI team. The meetings enabled researchers to determine analytic challenges in the organization's big-data-driven decision making environment. Active participation in the various analytic projects helped researchers pinpoint crucial gaps in literature and reinforced the need for an updated process model for KDDA. In the following eight months, researchers refined and evaluated a new process model through multiple build-and-evaluate loops.

4. Snail Shell KDDA process model

As discussed in Section 3, our artifact design starts with articulating business needs or "problems" that the new KDDA process model should address. The large volume, variety, and velocity of big data environment demands not only the scalability of analytic solutions and deployment, but also the faster turnaround from problem formulation to decision making. This requires more frequent iterations among different phases compared to traditional KDDM process models.

The snail shell model (Fig. 1) consists of eight phases with arrows indicating the highly iterative nature of the KDDA process. It assimilates the key concepts (phases, tasks, and guidelines) of the KDDA process at the meta-level, and inherits the project life cycle representation from CRISP-DM process model. There are no strictly defined sequences between phases, though most KDDA project starts with the problem formulation (PF) phase. Each phase includes different tasks, and the outcome of each task determines the phase or particular tasks of a phase to be performed. The arrows indicate frequent shifts between phases, where it starts with the output of one phase and directs to the phase to be performed next. The bi-directional arrow means the movement can be to and from the phase. For example, a specific output of the modeling phase may requires going back to business understanding, data understanding, data preparation, or going directly to the evaluation.

Compared to traditional KDDM models, two additional phases, namely problem formulation (PF) and maintenance, are introduced in the snail shell model. The quality of a well-formulated business problem can potentially affect the results of succeeding phases in the KDDA process. In traditional KDDM, the formulation of business objectives and data mining goals focuses on describing and solving well-defined problems with limited guidance for problem formulation. In the area of analytics, the business problems are often ill-structured and complex. Not only should KDDA understand the business problems, business users also need to have a realistic expectation of what analytic models can achieve. Therefore, it is important to introduce the PF phase to guide the systematic formulation of achievable analytic problem statements from stated business objectives. Furthermore, the big data environment makes the model maintenance more complex. Although traditional KDDM models embed model monitoring and maintenance in the deployment phase, it only focuses on planning-related activities, such as identifying potential model changes, describing how to monitor model performance, determining when to update or retire models, and documenting business problems. Model maintenance in KDDA covers more than just planning for changes in the business environment or data. It should track all aspects of the analytic model life cycle. Hence, we introduce a model maintenance phase. In the following sections, we outline each phase, and provide a comparison between respective KDDA and traditional KDDM tasks.

4.1. Problem formulation (PF)

The PF phase involves formulating the business problems that the KDDA project should address, and transforming it to an actionable analytic problem statement. A problem can be best defined as an undesirable situation that is solvable with some difficulty, and is expected to be altered or completed in a desired manner [1]. Problem formulation has been well recognized as the most important aspect of the decision making process [32,34]. Literature [34] identifies four types of problem

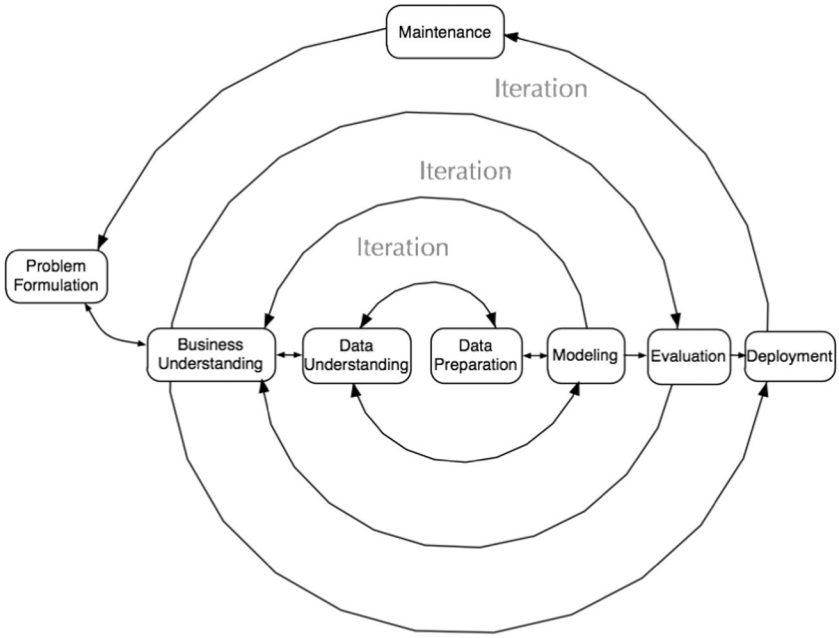

Fig. 1. The snail shell KDDA process model.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

formulation processes that are applicable to the PF phase. They relate to the: *clarity of the goal state*, *characteristics of its problem space*, *set of problem-relevant knowledge*, and *problem solving process*.

Literature suggests various problem formulation strategies that include formal problem representation as models [33], reformulation [17], decomposition (factoring complex problems to manageable small ones) [41], and heuristics [42]. Techniques such as Value Focused Thinking (VFT) [25], Goal Question Metric (GQM) [44], and SMART [15] provide some structure towards formulating business problems in the ill-structured decision context of KDDA. For example, GQM approach can be adopted to establish measurable goals, and SMART criteria (Specific, Measurable, Achievable, Relevant, and Time-bounded) can be used to assess organizational objectives. As a KDDA project moves through different stages, new problems may become apparent. Refining the problem statement can often become a necessity to ensure timely and achievable KDDA objectives. The iterative nature of the snail shell model provides a systematic way to identify and address needed revisions. Table 1 provides a summary of tasks for the PF phase and a comparison of KDDA with the traditional KDDM approach.

4.2. Business understanding (BU)

The BU phase focuses on the business requirement elicitation that ultimately helps to translate high-level executive requirements into specific analytic needs. One of the key tasks in this phase is enterprise knowledge acquisition that includes acquiring both tacit–explicit and individual–collective knowledge [35]. Explicit knowledge may reside in multiple sources, such as Enterprise Content Management (ECM) systems, ETL processes, and BI reports. Access to tacit knowledge is essential to acquire and interpret explicit knowledge. Similarly, shared knowledge should be accessed to obtain collective knowledge.

ACM assessment is another key task in BU. An organization's analytics capability may be assessed along three dimensions: *organizational analytics maturity* that describes the analytic environment in the organization; *data maturity* that defines whether data is suitable for analytics; and *decision style maturity* that describes whether the business users' decision styles are mature enough to use analytics results. A more comprehensive ACM model shall be developed to guide the analytic process improvement.

The *data maturity* assessment focuses on availability, stability, and quality of data for analytics. Data stored in the big data platform and EDW may not necessarily be ready for analytics. For example, Cassandra databases can be used to capture transactional big data (web click streams, machine event logs, sensor tracking data, etc.) Without adequate planning, data may not be suitable for analytics as they are not queryable. Similarly, NoSQL databases are designed to store schema-less data, but it does not mean that data modeling can be ignored. On the contrary, data modeling becomes more challenging in the big data environment, as it requires specific design considerations of the data access path.

One approach to assess *decision style maturity* is to use the dynamic decision style model [16] which includes two factors: *information use* (i.e. the amount of information considered in making a decision) and *focus* (i.e. the number of alternatives identified when reaching decisions). Two decision patterns are related to *information use*: the *satisfier* who tries to reach "good enough" situation based on just enough information, or the *maximizer* who wants to get all relevant information before making a decision. Two different decision patterns are related to *focus*: a person with *unifocus* lens uses information to produce one solution, and a person with *multifocus* lens uses several options. The decision style model combines the four dimensions and defines five decision styles, as shown in Fig. 2. Each decision style has its strengths and weaknesses. However, it is easier to present analytic results to the *maximizer* than to the *satisfier* because the latter trusts his or her instincts to find good enough solutions.

Existing KDDM process models tend to prefer traditional SDLC methodologies. The iterative nature of the KDDA process, however, requires thoughtful consideration to determine the appropriate project management (PM) methodology. Based on the nature of the project, the KDDA project team may choose waterfall, agile, or mixed methodology. Relevant tasks for the BU phase are summarized in Table 2 and a comparison with the traditional KDDM approach.

4.3. Data understanding (DU)

This phase involves familiarizing with data from various sources that are relevant to solving the analytic problem. Though traditional KDDM model prescribes that the analytics practitioner collects initial data before describing data, the increasing complexity and variety of today's big data environment make the initial data collection almost impossible without a thorough understanding of the existing data structure, size, and format. Furthermore, humans have known cognitive limitations when too much information is presented in an inappropriate way. Visual analytics is "*the science of analytic reasoning facilitated by interactive visual interfaces*" [43]. The integration of visualization and modeling can provide support for data understanding in the analytic process [26].

Since data quality is always a concern in analytics, appropriate data quality metrics have to be chosen based on business and analytic requirements. For example, time series analysis does not allow missing data for the time interval, which would not be an issue for decision tree analysis. Data description is another important task to understand initial characteristics of the data, its metadata, source systems, update frequencies, etc. Table 3 lists the set of tasks related to the DU phase.

4.4. Data preparation (DP)

Based on outputs from the PF, BU, and DU phases, an initial data integration requirement shall first be created, including how each data element for modeling would be sourced or transformed. Multiple iterations of the DU, DP, modeling, and BU may be performed until an acceptable modeling dataset is produced. All quality related issues

Table 1
Problem formulation tasks summary.

KDDA tasks	Description	KDDM comparison
Determine business objectives and success measures	Various techniques can be used to facilitate goals and objectives determination, including VFT, GQM, SMART, mean-end analysis, etc. Objectives shall be directly or indirectly measurable.	Similar.
Deploy problem formulation strategies	Determine boundaries, factor complex problems into sub-problems; focus on controllable components of a decision situation, reformulation, and heuristics.	Not available.
Define business problem	The business problem shall have a problem definition that are related to what, why, and how questions.	Not available.
Determine KDDA problem, goals, and success measures	The KDDA problem type and goal are determined based on the business problem and objectives. The analytic goal needs to be measurable with formally defined analytic success measures, which is the input for evaluating final analytic models.	Similar to tasks in the BU phase of KDDM.

	Satisfier	Maximizer	
Unifocus	Decisive	Hierarchic	Systematic
Multifocus	Flexible	Integrative	

Fig. 2. Five decision styles [16].

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

Table 2
BU summary tasks.

KDDA tasks	Description	KDDM comparison
Establish business case	Identify and quantify costs and benefits; identify requirement, assumptions and constraints, risks and contingencies; inventory of resources; present business case to executive sponsor.	Part of the situation assessment task, but with limited specifications.
Analytics capability maturity assessment	Includes three types of maturity assessment: data maturity assessment, organizational maturity assessment, and decision style assessment.	Not available.
Enterprise knowledge acquisition	Includes explicit knowledge acquisition from existing document, business process, ETL process, queries, BI reports, relevant matrices, data quality requirement and matrices, etc. Also includes tacit knowledge acquisition.	Not explicitly stated.
Determine PM methodology	Understand the nature of the project as well as the organizational culture.	Implicitly defined as iterative SDLC.
Initial tools and techniques selection	Software selection framework can provide some decision support. It is constrained by the business case output.	No software selection framework provided.

identified in the DU phase should be addressed, which may be handled within or outside of the analytic tool. The actual strategy is contingent upon the unique decision making situation (DMS) and quality requirements of selected analytic techniques. For example, regression analysis and decision tree analysis are both robust towards missing values, whereas null values must be replaced for time series analysis. Tasks in the DU phase are summarized in Table 4.

4.5. Modeling

This phase involves selecting applicable modeling techniques and building analytic models to provide most desirable outcomes for the stated analytic goal. The technique selection is constrained by tool(s) chosen in the BU phase. For example, One Class Support Vector Machine, a popular abnormality detection technique, is not available in many analytic packages, including SAS Enterprise Miner and SPSS modeler. Each modeling technique has its own process requirement, where modeling rules must be followed (e.g., imputation of missing values). After fine tuning parameters of each model, an initial assessment of modeling results needs to be conducted to identify the set of candidate models for further evaluation. Tasks related to the modeling phase are summarized in Table 5.

4.6. Evaluation

In this phase, candidate models are evaluated against business objectives and business problems formulated in the PF phase. While analytic models may be evaluated within the tool using objective measures such as accuracy or lift, evaluation against business objectives is often not clear-cut. Thus, test scenarios may be constructed and additional evaluation criteria may be defined. The initial model may be tested in a real-world application, but at a much smaller scale. The KDDA process shall be reviewed to determine if there are other factors or tasks that have been overlooked, and to understand how the modeling requirement may influence existing business, data, and analytic processes. For example, what kind of ETL change may be needed to deploy the model in production? Communicating results to executive sponsors and stakeholders is another task in this phase. Once the models and modeling process are reviewed and signed off, the next step is to determine if additional iterations are necessary prior to deployment. The evaluation result may reveal that the business problem has not been

Table 3
Data understanding summary tasks.

KDDA tasks	Description	KDDM comparison
Within-DBMS data exploration	SQL can be used in relational databases. Pig, Hive, or other NoSQL query languages can be used in big data platforms.	Only available after initial data collection.
Out-DBMS exploration	Advanced visualization tools are recommended.	Similar.
DU for business requirements	Many business requirements and business logic reside in the data or data related processes.	Not available.
DU for modeling requirements	Depending on the selected modeling technique, different types of DU need to be performed.	Not available.
Verify data quality	Data quality depends on business requirements, as well as the analytic techniques selected.	Similar.
Describe data	Data description should include source, owner, update frequency, and other relevant attributes.	Similar.

Table 4
DP summary tasks.

KDDA tasks	Description	KDDM comparison
Create data integration requirements	Based on BU and DU outputs, requirements for the modeling dataset (e.g. how to source each data element) shall be created and communicated. A data integration strategy should be defined (e.g., to integrate data on the fly or create a new ETL process).	Similar, but separated into two tasks: dataset and select data.
Data transformation based on quality requirements	Data cleaning and transformation are closely related. The same data quality issue may need different types of transformation based on selected analytic techniques	Similar to clean data task.
Data transformation based on business requirements	Transformation may be needed based on business requirements, such as normalization and aggregation.	Not available.
Data transformation based on modeling requirements	Depending on the selected modeling techniques, different types of data transformation may be needed Integrated knowledge repository with modeling rules for KDDA can provide decision support in this task.	Similar to construct data task.
Data integration	Integrate data based on formally defined and approved data integration requirement.	Similar.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

Table 5
Modeling summary tasks.

KDDA tasks	Description	KDDM comparison
Select modeling techniques	Based on the analytic problem and requirements in BU phase, suitable modeling techniques are selected.	Similar.
Describe modeling rules for the modeling technique	Each selected technique includes a set of modeling rules. An integrated KDDA knowledge repository can provide decision support in this task.	Not available.
Define training and testing strategy	Define how analytic models will be trained and how their performance can be assessed.	Similar.
Build models	Modeling rules shall guide the process of building model. Any additional insights shall be documented and used to update modeling rules in the future.	Similar.
Assess models	Models are assessed using previously defined criteria, and candidate models are chosen for further evaluation.	Similar.

adequately addressed, or that a new problem formulation is needed. Table 6 summarizes relevant tasks in the evaluation phase.

4.7. Deployment

The strategy for deployment shall be considered early in the KDDA process as part of the BU phase. It is important to share the deployment plan with all stakeholders and ensure that resources are available. Similar to the traditional KDDM process, a deployment plan is one output of this phase. It should summarize the deployment strategy, and outline steps to perform them. The deployment plan shall also document all nonfunctional requirements (e.g., security and performance requirements) and functional requirements (e.g., systems, database, and network infrastructure). All changes needed for the analytics process in the organizational environment shall be documented and communicated with stakeholders for their buy-in. Tasks for deployment are summarized in Table 7 below.

4.8. Maintenance

The maintenance phase includes all model management activities, including model selection, usage, retirement, and replacement. A formal model maintenance process needs to be established with assigned roles for business users. Clearly defined guidelines and procedures are needed to specify when and how to implement changes in the deployed analytic models. Model changes may be necessitated when there is performance deterioration, or changes in the data or business environment. Based on the functional and nonfunctional requirements documented in the deployment phase, model usage shall be monitored and feedback from users shall be collected. Collective performance of models sets also needs to be monitored [29]. Table 8 summarizes tasks in the maintenance phase.

5. Device abnormality behavior detection case study

Similar to Chiang & Che [9], we use a device abnormality behavior detection project to evaluate our proposed model from both understanding- and action-oriented perspectives. Due to space limitations, we do not present all four KDDA projects that were guided by the snail shell model over the duration of eight months. Instead, we elaborate one full cycle of the first KDDA project that was successfully closed within 10 weeks. While it is impractical to present all the transitions between phases, especially during the initial PF, DU, and DP phases, we describe main tasks in each phase of the project and major iterations between phases to demonstrate the effectiveness of implementing the snail shell process model. Due to organizational policy, all identifiable information is either removed or anonymized.

5.1. Problem formulation

When the first big data analytics project was initiated, the director has not yet had a clearly articulated business problem at hand. The main business objective was vaguely described as "*find out from data what went wrong in the SBO environment.*" The executive director acknowledged the complexity of the decision environment as: "*SBO environment has many variables that impact its performance — hardware, software services, and even weather. We collect so much data on so many different dimensions that we are not able to determine what change(s) that may cause the performance deterioration…*"

Guided by the PF phase of the KDDA model, the researchers started with defining boundaries of the problem by raising pointed questions of *What*, *When*, *Where*, *How*, and *Who*? Combined with enterprise knowledge, this exercise enabled researchers to formulate two problem statements in a more declarative form: (1) "*What has changed in the SBO environment that causes unwanted reactive events?*", and (2) "*How can changes be determined to take proactive corrective actions?*" Several problem formulation strategies were deployed, that included focusing on controllable components of the decision situation and setting limits on the problem boundary to two directly obtainable device performance measures: error logs and reconnect logs. This helped structure the problem statement more specifically as: "*How can error and reconnect logs be used to identify these needles (SBO device with issues)?*"

The statement itself lacked the clarity and actionability, as there were no clearly defined business objectives. The director ascertained that the high-level business objective was to reduce resources spent in diagnosing past events. Researchers therefore decided to utilize the GQM approach to formulate the business objective, and subsequently evaluated using SMART criteria. This thorough and iterative analysis refined the business objective as the need to "*identify SBO devices that are abnormal from their usual state near real time so that the analysts can focus on investigating these devices (the needles) rather than querying the whole device pool (the haystack)*".

Table 6
Evaluation summary tasks.

KDDA tasks	Description	KDDM comparison
Evaluate result	Evaluate the models based on specified business objectives and requirements. If direct evaluation is not possible then a field test may be required.	Similar.
Conduct field test	Create test cases and test the model in a testing environment in situations where business objectives or requirements cannot be directly evaluated.	Not available.
Review analytic process	It shall include questions related to whether additional insights are beneficial to the organization, whether changes are needed in the business process, data process, or analytic process, etc.	Similar.
Communicate results	Analytic results shall be communicated effectively with executive sponsors and stakeholders, and refer back to the business case presented in the BU phase.	Not available.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

Table 7
Deployment summary tasks.

KDDA tasks	Description	KDDM comparison
Create deployment plan	The deployment plan shall explicitly document functional and nonfunctional requirements.	Similar, but does not highlight requirement documentation.
Produce final project report and final presentation	Each project iteration shall be well documented. Presentation of results to management is critical for successful deployment and future maintenance.	Similar.
Review Project	Any additional modeling rules about the modeling process shall be documented and stored in the KDDA knowledge repository.	Similar, but does not call for recording discovered rules about the modeling process.

Based on the identified business objective, the analytic problem was synthesized as "estimating a device's current state as being *normal* or *abnormal*". Communicating PF tasks with the director helped frame the following KDDA modeling requirements: 1) analytic models should be easy to build and deploy; 2) they should be stable (i.e. require minimal human intervention); and 3) the KDDA process should be flexible so that it can be expanded to include other relevant dimensions.

5.2. Business understanding (BU)

In this section, we describe two BU iterations related to the analysis of error and reconnect logs, and challenges that were factored into the problem statement at a later stage in the project. Immediately after the PF phase, the task of enterprise knowledge acquisition was carried out starting with a thorough review of the company's central knowledge base, an ECM system, to identify SBO related environment variables and their definitions. ETL processes, Splunk and Pig queries, and BI report were also reviewed to extract relevant business logic. The researchers held multiple conversations with business users to identify the combination of SBO hardware and software environments that were most relevant to the business problem. The BI analysts helped in interpreting error and reconnect logs. Tasks in the BU phase guided researchers to effectively combine both explicit and implicit knowledge in the problem domain. This also resulted in the creation of an improved team "onboarding" document for future team members.

A formal ACM assessment was then performed to determine whether the organization was mature enough to carry out a KDDA project of this scope. The organization positions itself as a technologically focused company with a deep data-driven decision-making culture starting from the top executives to the different business unit directors. However, the SBO BI team was still in the initial stage of descriptive analytics with no formal analytic processes in place to drive KDDA. The team also lacked an understanding of advanced analytics tools and techniques. From the organization maturity perspective, the analytics capability of the SBO business unit was determined as *Low* to *Medium*.

From the data maturity viewpoint, the organization was in the process of capturing and integrating machine data for analytics purposes. Log files were loaded from Cassandra to Hadoop Distributed File System (HDFS) through Apache Flume, and indexed by Enterprise Splunk. Apache Pig scripts over MapReduce encoded the data flow and guided the HDFS schema design. Furthermore, HDFS log files were loaded into and integrated with the Enterprise Data Warehouse (EDW) through an automated ETL process. Disparity in data stored and managed in these multiple sources created enormous data integration challenges. The data maturity was thus assessed as *Low* for analytics. From the decision style maturity perspective, the director was a mix of *decisive* and *systemic*. In situations where there was time pressure to make quick decisions on less complex problems, he was able to make speedy, efficient, and prompt decisions by settling on 'good enough' information. In more complex situations, he would solve the problem in a *systemic* style. Decision styles of other leaders were also very similar. Based on the assessment, researchers determined that the organization had the ability to answer, "*What has happened*", as posited in the PF phase. The ACM assessment was communicated with the top management, and the company initiated steps to improve its organization and data maturities.

Another task in this phase was to determine the project management methodology. As expressed by the director, "*we are not a process-oriented team*". Based on the team culture and the nature of the problem, the researchers and the team agreed on a semi-agile methodology, and consequently, the time-box iteration concept [2] that involved four iterations of task planning, development, demo and retrospective was adopted. The time-box was kept at one week and tasks were adjusted accordingly.

To determine if error logs and reconnect logs could be used to answer analytic questions, a second iteration of BU was initiated. It started with conversations with BI analysts who understood available data sources and structures. This round of BU iteration, paired with the business requirement task of the DU phase, factored two specific sub-problems: (1) *How can error logs be used to detect the SBO environmental changes?* and (2) *How can reconnect logs be used to detect SBO environmental changes?* For a deeper assessment of whether error logs and reconnect logs data could be used to answer these questions, multiple iterations between BU, DU and PF were performed.

Three main data sources were accessed to obtain error logs and reconnect logs. First, Cassandra stored the most recent one-month logs (indexed by enterprise Splunk) in a comprehensive manner. Second, HDFS stored the past one and half years of error logs and reconnect logs with certain information removed because of the limited storage capacity. Third, EDW stored one-month error logs and reconnect logs integrated with non-machine generated data. The one-month truncated period was chosen as the organization's business reports only

Table 8
Maintenance summary tasks.

KDDA tasks	Description	KDDM comparison
Describe and store analytic results	Analytic models are semantically described, including its data input and transformation, modeling technique and parameters, model performance measures, and business performance measures.	Not available.
Create a model maintenance process	The process shall include initiating maintenance cycle, delivering model result, checking model performance, preparing change report, authorizing and performing model update, and communicating updates with business users. Each activity in the process execution shall include roles with accountabilities.	Not available.
Define change initiation	Explicitly describe how to capture changes related to model performance, business environment, and data.	Not available.
Monitor model usage	The model usage shall fit its security requirement. End users' feedback on the model usage shall be included, which may initiate additional changes not formally defined in the change initiation.	Not available.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

concerned most recent history. In addition to exploring stored log files, researchers also reviewed organizational dashboards, BI team reports, and ECM documentation to gain a better understanding of different types of error and reconnect codes. Six reconnect codes were identified and selected as the focus of the study through this iterative knowledge acquisition task. These set of considerations led to the reformulation of the problem statement as: "*How can SBO environment change be identified using the six types of reconnect codes?*" The next section describes two iterations of the DU phase to assess the error logs and reconnect logs. It also provided the input for the data integration plan later in the KDDA process. The steps further demonstrate the iterative nature of the KDDA process.

5.3. Data understanding

As mentioned previously, the BU phase established that the ultimate goal was to identify devices with abnormalities so that the BI team could focus on investigating these devices using only dimensions that generated the issues (*the needles*), rather than looking at everything (*the haystack*). Any intermediate results were communicated with the BI team to interpret its meaning and decide next steps. Two main objectives for DU were identified: 1) *Assess whether error logs were meaningful to identify changes in the SBO environment* (e.g., a software service update might result in a higher number of failures related to one type of error codes), and 2) *Assess whether reconnect logs were meaningful to identify changes in the SBO environment* (e.g., the device location and hardware might validate reconnect reasons). Next, we describe activities related to each objective.

The researchers focused on error logs in the first DU iteration. Within-DBMS data exploration was the first task performed in this phase. There were more than one hundred different error codes, and as mentioned before, the data sources for the logs differed in their level of granularity and historical load intervals. Researchers wrote Splunk queries to review statistics and indexes related to error logs from Cassandra datastore. However, as error codes were summarized by services and software versions, Splunk was not sufficient to provide error logs at the device level for analytics.

Since performance was a concern when querying HDFS, researchers used a closed date range to limit the number of concurrent MapReduce jobs, and to reduce the associated query execution time. To gain a comprehensive view of data, eight months of error logs were examined one month at a time. Within-DBMS exploration revealed many data abnormalities. For example, the distribution of error codes identified two event abnormalities, which were addressed subsequently in the DP phase. Another observed abnormality was that one error string did not have an associated error code. After returning to the BU phase to better understand the abnormal event, researchers were able to establish that the null error code was misreported and not a concern. The process exemplifies the significance of the snail shell model's iterative approach at every step of the KDDA process. Splunk was the only place where software-related information was formally captured. However, only one-month history was available in Splunk, which was not sufficient for analytic purposes. These findings indicated that the related analytic question could not be answered without further integration of error log data from other sources. A data integration plan was created to correlate software services with error logs, as discussed in Section 5.4.

The second DU iteration focused on understanding reconnect logs. Researchers wrote Hive queries to understand the distribution of different types of reconnects. The queries were guided by the ETL business logic to ensure consistency. For example, although multiple applications generated reconnect logs, only reconnect codes related to "core" applications were used for the analysis, 99% of which consisted of six reconnect reason codes. The findings reconfirmed the revised problem statement in BU iteration 2. Both DU iterations highlighted importance of thorough data exploration in its original format. As commented by the director, "…*only when modeling results show added business value that this process could be formalized in the ETL process*."

Visualization is an important analytic technique [36] that facilitates out-DBMS exploration. Tableau visualization revealed two different periods where reconnects were much higher in the first period then the second period. The BI team indicated that the change was the result of recent SBO application process improvements. Hence, data were truncated to include only the stable period (most recent 10 weeks). Visualization also discovered potential data quality issues. For example, two subregions *South* and *North* were sourced in the system in two different forms: *South* and *South Region*; and *North* and *North Region*. The data quality issue was reported to the ETL team. As an ETL process change to correct this error would take time to implement, the researchers created a process in R to solve this issue.

For the next task, i.e., DU for modeling requirement, data was loaded to R through RJDBC connection and R code was written to check data distributions and modeling assumptions. Box plots (Fig. 3a) were used to inspect if there were differences between group means for dimensions. Density plots (Fig. 3b) were used to test density by groups.

The DU outputs indicated a close-to-normal distribution of reconnects. The central limit theorem states that the sampling distribution of any statistic will be normal or nearly normal, if the sample size is large enough. The rule of thumb is that the sample size should be greater than 40 without outliers. The sample size for the reconnect analysis was 70 days and thus fits the central limit theorem. Outliers and missing values were handled using R code during the data preparation. R *sqldf* function was used to find distinct class variable instances and then calculate expected observations. The final integrated training dataset had 134,485 rows.

5.4. Data preparation

In this section, we present three iterations of the DP phase. The first iteration was to prepare data for analyzing error logs. Although an initial integrated dataset for error logs was created, the team agreed that utilizing the error logs in a meaningful way would require significant

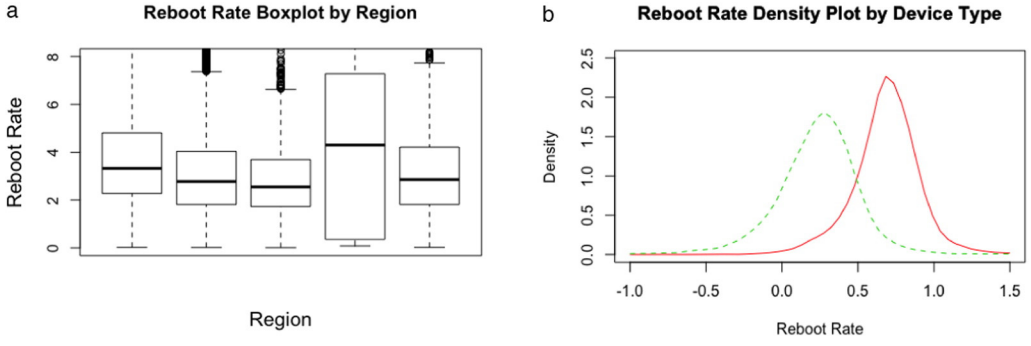

Fig. 3. a: Example of box plot to compare group mean by region. b: Example of density plot by device type.

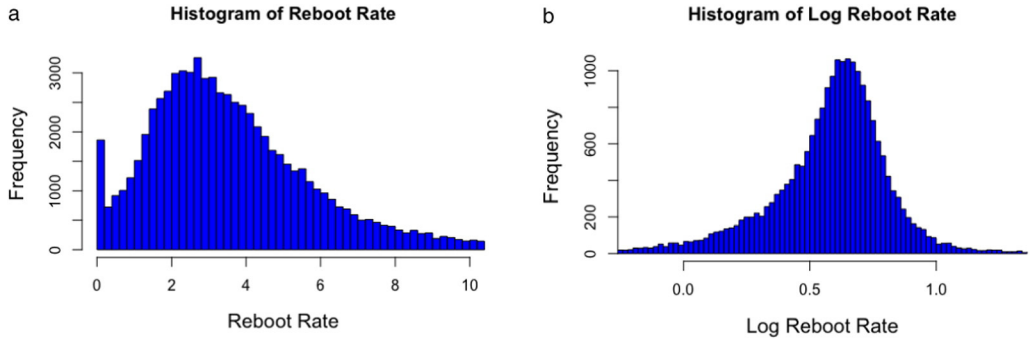

Fig. 4. a: Histogram of reconnect rate before transformation. b: Histogram of reconnect rate after log transformation.

changes to the existing ETL process. The second and third iterations were to prepare data for analyzing reconnect logs. Six integrated datasets were created and transformed, one for each reconnect reason code.

The researchers first started with a data integration plan to collect device-level error logs. As part of DP tasks, an initial data integration requirement for error logs (e.g. how each data element should be sourced from different data sources) was created and communicated with the BI team. The first step was to correlate software versions with device types based on change management files from the ECM. Most recent 16 weeks of error logs in HDFS were retrieved and ingested in EDW through SQL Loader. Multiple queries were written to join error logs with related SBO dimensions. When the result was presented to the director and BI team, the director commented: "*It (data integration process) is not sustainable with our limited resources.*" The BI team also indicated that, even if the modeling result could detect an increase in error rate due to a change in the input variables, no actions could be taken to peruse the result as error codes were collected in different granularity levels. Therefore, the director decided to delegate the analysis of error logs to another KDDA project in the future.

The output from BU iteration 2 and DU iteration 2 guided the researchers in producing a data integration plan for reconnect logs. The data transformation task was carried out in three steps: 1) retrieve reconnect logs from HDFS, 2) load data to EDW, and 3) add device related dimensions. To address the data quality issue related to the subregion names discussed in DU iteration 2, a data transformation process was designed and implemented in R. Additional data transformation tasks were also performed based on the business requirement.

The third iteration undertaken was a data transformation task based on the modeling requirement. For example, the histogram of the reconnect rate (number of reconnects per 1000 devices) for reason code 1 showed a skew normal distribution (Fig. 4a). To fit the normality assumption of the modeling requirement, the skewness was mitigated by log transformation of the reconnect rate (Fig. 4b).

Two strategies were considered for outliers: replacement and removal. R packages for finding and replacing extreme values were tested on the sample dataset, and none of them performed well. Therefore, researchers created a removal strategy based on the unique modeling requirement. Customized R code was written to remove outliers. The tasks were documented, and the results were presented to the director and BI team before moving to the modeling phase.

5.5. Modeling

We present two important modeling iterations in this section. As discussed earlier, the findings from the DU phase indicated that error logs should be excluded from the scope of the project. Therefore, the modeling phase only concerned reconnect logs. Using the six integrated dataset from the DP phase, each reconnect reason code was modeled individually.

The first iteration started with selecting suitable modeling techniques. The target variable, reconnect rate, in the final dataset was continuous. All directed learning methods for interval targets (e.g. regression, regression tree, artificial neural network, and K-nearest neighbor) were considered as relevant techniques. However, since the goal was to provide the BI team with the ability to analytically diagnose reasons for device abnormality, only explanatory modeling techniques were applicable, which led to the initial selection of regression and regression tree techniques.

Regression analysis was the first modeling technique tested, with reconnect rate outside the 95% confidence interval considered abnormal. However, the first regression run did not produce significant results, and the team agreed that the estimation of 95% confidence intervals was too narrow with too many false negatives. A traditional control chart approach was recommended by researchers to discover reconnect variations using the group mean and standard deviation. The UC (Upper Control) limit was set as two standard deviations above the mean, and LC (Lower Control) limit was set to zero.

The second iteration targeted finding the aggregated group mean. By the time this iteration started, two additional weeks of data were available for testing purposes, and consequently all 70 days of data were used in finding group mean and standard deviations using the R *aggregate* function. The R code for the model is shown in Fig. 5.

```
GroupMean<-aggregate(Log_Reboot~MANUFACTURER_TYPE+REGION_NAME+HOUR_KEY,
RECONNECT, mean)
colnames(x)[4]<-"Mean"

GroupSd<-aggregate(Log_Reboot~MANUFACTURER_TYPE+REGION_NAME+HOUR_KEY, RECONNECT,
Sd)
colnames(y)[4]<-"Sd"

GroupResult<-merge(GroupMean,GroupSd)
GroupResult<-merge(GroupResult, Market, all.x=TRUE)
```

Fig. 5. R code for control chart modeling.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

Table 9
Outcomes of adopting the snail shell KDDA process model.

Phase: task	Issues before model adoption	Outcomes from model adoption
PF: clearly articulated business problem statement.	The business objective was vague and the team was acting in a reactive mode. They considered themselves "*pretty good at understanding what has happened*."	Clear definition of business problems and measurable analytic goals helped set realistic expectations from the top management. Refining problem statement was justifiable through this formal process.
BU: enterprise Knowledge Acquisition	Lack of holistic knowledge of the problem space made it a lengthy process to "onboard" new team members.	Comprehensive examination of organizational knowledge base improved efficiencies in BU and DU, especially when eliciting tacit knowledge.
BU: formal ACM assessment	Organization did not know if it could use data and analytics to solve business problems.	A formal ACM assessment not only helped establish project feasibility, but also provided recommendations for improving analytics capability from organizational, data, and people perspectives.
BU: project management methodology	Traditional SDLC life cycle for KDDM was not suitable for the team culture and for the project.	Adopting semi-agile methodology allowed team collaboration on the KDDA project and shortened project life cycle.
DU: data exploration within its original format	Data exploration after initial data collection was not practical in the big data environment.	Quick profiling of big data in its original format (i.e. machine logs) not only enabled discovery of interesting data elements and data abnormalities, but also avoided additional ETL workloads on data engineers.
Evaluation: test model in the field	Using historical data with clear evaluation measures (e.g. accuracy) was not practical in some cases.	Field test helped evaluate modeling results that were otherwise not directly measurable.
Model management	No guidelines to manage model life cycle; model effectiveness often impacted by frequent changes in the big data environment.	Model maintenance enabled faster detection of model performance deterioration and more frequent model updates.

Since the trend was not captured in this approach, researchers recommended an automated process to refresh group means and standard deviations using new incoming data. The modeling results were assessed based on the requirements specified by the director during the PF phase (i.e., easy to build and deploy, stable, and flexible). The first requirement was satisfied, as the snail shell model offered the advantage of providing a systematic approach to address all activities in model development. The second requirement was contingently satisfied. The control chart approach was robust and would remain so unless there were radical changes to the SBO environment. The third requirement was achieved as the model presented a repeatable process for the SBO environment. As long as the data normality assumption is satisfied, additional dimensions may be added.

5.6. Evaluation

Each modeling result was evaluated using the most recent two weeks of data. The director designated a BI team member to inspect the subregions that had abnormally high reconnect rates. Since there was no single indicator for the abnormal behavior, a direct evaluation of the modeling result was not feasible. Therefore, a field test approach was selected. A detailed review was initiated to investigate correlations between modeling results and SBO environment changes. For example, when one subregion was identified with abnormally high reconnect reason code 2 for consecutive hours, the BI team contacted the subregion's device quality team, and found out that, while there were no internal SBO environment changes, the sub-region was hit by a major storm during these hours. This finding highlights the complexity of SBO environment, where many variables could impact the device performance (e.g., hardware, software, customer's home characteristics, and even weather.) The model helped to identify these *needles* in the *haystack* that otherwise would have remained undetected.

Additionally, in reviewing the analytic process, researchers determined that one-month reconnect logs currently stored in EDW were not statistically sufficient for data analytics, and so recommended that the ETL process be changed to include 90 days of reconnect logs. This change also made model maintenance easier. For completing the task of communicating results, the analytic process was thoroughly documented and a wrap-up presentation was given to the BI team and other business users. The modeling results sparked an interest within the BI team to investigate additional reconnect reasons. The question emerged, "*can we use reconnects data, or some other machine data, to identify a device in an unhealthy state?*" This problem statement kick-started a follow-up analytic project. This further demonstrates the significance of the iterative nature of the KDDA process.

5.7. Deployment

Developing the deployment plan was relatively straightforward because the modeling technique aimed at easy deployment. A modeling table was created to store means and standard deviations for all six types of reconnect codes, and a deployment SQL statement was written to run against EDW. The modeling results were also stored in the EDW so that the reconnect behavior history can be tracked. Interesting findings related to the KDDA process were documented and shared in the organization's ECM.

5.8. Maintenance

As analytic initiatives were new at the organization, no centralized model repository was available. Once the organization improves its analytics maturity level, the model shall be semantically described and stored in a centralized model repository for reuse. A model maintenance process based on the modeling approach was designed and started immediately following the initial deployment. An hourly reconnect control chart was created in Tableau. If more than 10% of subregions were identified with abnormal reconnect behaviors, the model assumptions would require re-examination. The process also included a weekly refresh of group means and standard deviations to capture the trend. The models were implemented in the EDW with inherited security measures, and only used by the SBO BI team.

Less than three weeks after the deployment, one of reconnect codes (*reconnect reason* 2) regressed from the control chart, while another (*reconnect reason* 1) alerted more than 50% of subregions with abnormalities. The BI team was able to quickly trace the cause (application developers had modified one process to redirect reconnect reason calls). From a business point of view, reconnect reasons 1 and 2 represent two different types of user experiences. They were two leading metrics for measuring SBO device health. The issue highlights challenges in the today's big data environment, where the lack of appropriate data governance leaves the control of data quality and integrity at the hands of application developers.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

5.9. Case study summary

Throughout the project life cycle, the snail shell process model not only served as a roadmap to analytics experts, but also helped the BI team understand analytic processes and best practices associated with KDDA. Before adopting the formal KDDA process, the BI team mainly operated in a reactive mode, where intensive efforts were spent in analyzing historic data to understand "*What has happened?*" As remarked by the director, "*we're stuck in descriptive and edging into diagnostic*". The benefit of adopting the snail shell model was evidenced by the more "Why" and "How" questions raised during daily standup meetings. Table 9 summarizes some important decision support outcomes from adopting the snail shell model in the organization's analytic process.

6. Research implications and limitations

This research has implications for both theory and practice. From a theoretical perspective, our study introduced an updated process model for KDDA. While recent studies have started to investigate the use of big data analytics to enhance business performance, no studies have comprehensively addressed KDDA as an end-to-end process in the big data environment. Our research thus sheds light on a crucial gap in literature in the emerging area of research on big data analytics. Specifically, we utilized real-world cases to illustrate: 1) the need for problem formulation phase for establishing realistic expectations of analytic outcomes; 2) the need for model management phase to monitor, update and/or retire models in a timely manner; and 3) flexibility to move between phases during the KDDA process. The snail shell model we propose in this research is based on existing KDDM process models [31,40], and may be viewed as an initial step towards filling the gap. By appropriately applying existing foundations and methodologies from the knowledge bases, our research achieves rigor, while simultaneously strengthening the foundations of scientific theory upon which it is developed and built. More notably, our study illustrates how design science research benefits from the rich body of literature, and how it reciprocates by contributing to scientific knowledge.

The study also has practical implications. Our proposed snail shell model emphasizes the goal of knowledge discovery in modern enterprises where the need for timely action from big-data driven analytics is inexorably intertwined with productivity goals. Using real-world cases of big data analytics projects, we demonstrate how the snail shell model supports different decision makers collectively to achieve analytic goals. Additionally, for practitioners undertaking organizational KDDA initiatives, the case study may serve as a useful guide to gauge factors crucial to the successful knowledge discovery. For example, a comprehensive ACM assessment can directly influence the feasibility of organization's analytic goals. The case study also demonstrated that, not having a clearly defined data team or architecture impacted the ability to deliver timely and valuable knowledge to information consumers. Similarly, the nature of big data applications placed data modeling at the hand of application developers. Modifying application code or processes without consulting business users could compromise the integrity of data that is essential for data-driven decision making. Useful knowledge lost due to inadequate communication and lack of data governance thus represents one of the biggest challenges in utilizing big data for analytics. Hence, prior to initiating an analytics project, managers should conduct a thorough assessment to address data quality, data integrity, and data governance. It may be beneficial to establish a diverse KDDA team that comprises of analytics experts, data architect, data engineers, and business users. Managers may also need to make sure that some agile structure is in place for the KDDA team to draw upon their daily analytic activities.

The study is not without limitations. One limitation is that the evaluation was limited to one division within an organization, which implies that it only reflected analytic initiatives in that division. However, it is almost impossible to fully account for all situations encountered within an organization, especially the one with the scale and size where we conducted the study. Future research to understand how additional factors (e.g., service quality, risk and opportunity constructs) may play a role in organizational knowledge discovery and their impact on the KDDA process model would be worthwhile. A concern attributable to case study based evaluation pertains to its methodological rigor, researcher subjectivity, and external validity [2,47]. Although an unavoidable criticism, we provide explicit and ample details in the case study evaluation such that the approach is transparent, and other researchers may draw their own conclusions. We therefore view the evaluation approach not as a limitation. Instead it provides validation of the snail shell model in driving organizational KDDA projects. Nevertheless, quantitatively assessing the model using a large-N sample would be beneficial to evaluate its general suitability in supporting big data analytics in other organizational settings, and we encourage future research along this direction.

Another limitation of our research is that, we only explored three high level analytics maturity areas: organization, data, and decision style. Future research will focus on identifying additional analytics maturity areas, and developing best practices and metrics for each area. Future research may also investigate the development of a more comprehensive ACM model for analytics process improvement that includes other analytics maturity levels. For example, Gartner [22] recently published five maturity levels as unaware, opportunistic, standards, enterprise, and transformative.

The decision support domain for organizational big data is prime for additional research to establish standards of data quality, data integrity, and data governance for effective knowledge discovery. Similarly, agile analytics [12] (based on agile methods from software development) is gaining significant popularity in data mining and data warehousing. The unique characteristics of big data environments require careful consideration when adapting agile methods for analytics. Finally, research has not sufficiently addressed enterprise data architecture for decision support in big data environment, which represents another area for future research.

7. Conclusion

The popularity of big data and analytics has attracted the attention of researchers and practitioners. Recent changes, especially the proliferation of big data in business practices and the need for fast turnaround time from problem formulation to decision making, limits the effectiveness of traditional KDDM models for KDDA projects. While significant progress has been made in practical solutions for big data analytics, an updated process model for decision support in the big data environment is still missing. In this research, we propose a snail shell KDDA process model to address some challenges unique to the big data analytics environment. The design process of the snail shell model assimilates both existing KDDM knowledge bases and researchers' real-world experience in analytics. The utility and relevance of the proposed model was evaluated by scaffolding real-world at a global mass media company. For practitioners, the snail shell model may provide decision support for various decision makers in the systematic process of KDDA. It offers valuable insights for the parsimonious alignment of analytic projects and successful outcomes. For researchers, the study fills a gap in scholarship in addressing the limitations of existing KDDM models. It provides the starting point for further refinement of KDDA process knowledge.

References

[1] G.P. Agre, The concept of problem, Educational Studies 13 (1982) 121–142.
[2] A. Bennett, C. Elman, Qualitative research: recent developments in case study methods, Annual Review of Political Science 9 (2006) 455–476.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003

[3] A.G. Buchner, M.D. Mulvenna, S.S. Anand, J.G. Hughes, An internet-enabled knowledge discovery process, Proceedings of the 9th International Database Conference, Hong Kong July, 1999, pp. 13–27.
[4] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi, Discovering Data Mining: From Concept to Implementation, Prentice-Hall, Inc., New Jersey, 1998.
[5] R.M. Chang, R.J. Kauffman, Y. Kwon, Understanding the paradigm shift to computational social science in the presence of big data, Decision Support Systems 63 (2014) 67–80.
[6] S. Chaudhuri, U. Dayal, V. Narasayya, An overview of business intelligence technology, Communications of the ACM 54 (2011) 88–98.
[7] C.P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Information Sciences 275 (2014) 314–347.
[8] R.H. Chiang, P. Goes, E.A. Stohr, Business intelligence and analytics education, and program development: a unique opportunity for the information systems discipline, ACM Transactions on Management Information Systems (TMIS) 3 (2012) 1–13.
[9] T.-A. Chiang, Z. Che, A decision-making methodology for low-carbon electronic product design, Decision Support Systems 71 (2015) 1–13.
[10] K.J. Cios, L.A. Kurgan, Trends in data mining and knowledge discovery, in: L.C. Pal, N. Jain (Eds.), Advanced Techniques in Knowledge Discovery and Data Mining, Springer-Verlag, London 2005, pp. 1–26.
[11] C. Colburn, N. Towery, Introducing Surus and ScorePMML the Netflix Tech Blog, http://techblog.netflix.com/2015/01/introducing-surus-and-scorepmml.html2015.
[12] K. Collier, Agile Analytics: A Value-Driven Approach to Business Intelligence and Data Warehousing, Addison-Wesley, Boston, MA, 2011.
[13] H. Demirkan, D. Delen, Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud, Decision Support Systems 55 (2013) 412–421.
[14] V. Dhar, Data science and prediction, Communications of the ACM 56 (2013) 64–73.
[15] G.T. Doran, There's a SMART way to write management's goals and objectives, Management Review 70 (1981) 35–36.
[16] M.J. Driver, K.R. Brousseau, P.L. Hunsaker, The Dynamic Decision Maker: five Decision Styles for Executive and Business Success, Wiley, Hoboken, NJ, 1993.
[17] K. Duncker, L.S. Lees, On problem-solving, Psychological Monographs 58 (1945) i–113.
[18] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM 39 (1996) 27–34.
[19] S. Gregor, A.R. Hevner, Positioning and presenting design science research for maximum impact, MIS Quarterly 37 (2013) 337–355.
[20] G. Hahn, J. Packowski, A perspective on applications of in-memory analytics in supply chain management, Decision Support Systems 76 (2015) 45–52.
[21] A.R. Hevner, S.T. March, J. Park, Design science in information systems research, MIS Quarterly 28 (2004) 75–105.
[22] C. Howson, IT Score Overview for BI and Analytics, Gartner Group, Stamford, CT, Sep 2015 (Report No. G00291817).
[23] H. Hu, Y. Wen, T.-S. Chua, X. Li, Toward scalable systems for big data analytics: a technology tutorial, IEEE Access 2 (2014) 652–687.
[24] K. Kambatla, G. Kollias, V. Kumar, A. Grama, Trends in big data analytics, Journal of Parallel and Distributed Computing 74 (2014) 2561–2573.
[25] R.L. Keeney, Value-Focused Thinking: A Path to Creative Decision making, Harvard University Press, Cambridge, MA, 2009.
[26] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melançon, Visual analytics: definition, process, and challenges, in: A. Kerren, J.T. Stasko, J.-D. Fekete, C. North (Eds.), Information Visualization — Human-Centered Issues and Perspectives, Springer, Berlin 2008, pp. 154–175.
[27] A. Kumar, F. Niu, C. Ré, Hazy: making it easier to build and maintain big-data analytics, Communications of the ACM 56 (2013) 40–49.
[28] L.A. Kurgan, P. Musilek, A survey of knowledge discovery and data mining process models, The Knowledge Engineering Review 21 (2006) 1–24.
[29] B. Liu, A. Tuzhilin, Managing large collections of data mining models, Communications of the ACM 51 (2008) 85–89.
[30] Ó. Marbán, G. Mariscal, E. Menasalvas, J. Segovia, An engineering approach to data mining projects, in: H. Yin, P. Tino, E. Corchado, W. Byrne, X. Yao (Eds.), Intelligent Data Engineering and Automated Learning — IDEAL 2007, Springer, Berlin, Heidelberg 2007, pp. 578–588.
[31] G. Mariscal, Ó. Marbán, C. Fernández, A survey of data mining and knowledge discovery process models and methodologies, The Knowledge Engineering Review 25 (2010) 137.
[32] H. Mintzberg, D. Raisinghani, A. Theoret, The structure of "unstructured" decision processes, Administrative Science Quarterly 21 (1976) 246–275.
[33] W.T. Morris, On the art of modeling, Management Science 13 (1967) B707–B717.
[34] A. Newell, H.A. Simon, Human Problem Solving, Prentice-Hall Englewood Cliffs, NJ, 1972.
[35] I. Nonaka, H. Takeuchi, The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation, Oxford University Press, New York, 1995.
[36] O. Ola, K. Sedig, The challenge of big data in public health: an opportunity for visual analytics, Online Journal of Public Health Informatics 5 (2014) 223.
[37] F. Provost, T. Fawcett, Data science and its relationship to big data and data-driven decision making, Big Data 1 (2013) 51–59.
[38] M. Salehan, D.J. Kim, Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics, Decision Support Systems 81 (2016) 30–40.
[39] R.L. Sallam, D.W. Cearley, Advanced Analytics: Predictive, Collaborative and Pervasive, Gartner Group, Stamford, CT, Feb 2012 (Report No. G00230321).
[40] C. Shearer, The CRISP-DM model: the new blueprint for data mining, Journal of Data Warehousing 5 (2000) 13–22.
[41] H.A. Simon, The structure of ill structured problems, Artificial Intelligence 4 (1973) 181–201.
[42] G.F. Smith, Towards a heuristic theory of problem structuring, Management Science 34 (1988) 1489–1506.
[43] J.J. Thomas, K.A. Cook, A visual analytics agenda, Computer Graphics and Applications, IEEE 26 (2006) 10–13.
[44] R. van Solingen, V. Basili, G. Caldiera, H.D. Rombach, Goal Question Metric (GQM) Approach, Encyclopedia of Software Engineering, John Wiley & Sons, Inc., New York, 2002 528–532.
[45] S. Viaene, A. Van den Bunder, The secrets to managing business analytics projects, MIT Sloan Management Review 53 (2011) 65–69.
[46] Z. Xu, G.L. Frankwick, E. Ramirez, Effects of big data analytics and traditional marketing analytics on new product success: a knowledge fusion perspective, Journal of Business Research 69 (2016) 1562–1566.
[47] R.K. Yin, Case Study Research: Design and Methods, fifth ed. Sage Publications, Thousand Oaks, 2003.

Yan Li is an Assistant professor at Center for Information Systems and Technology, Claremont Graduate University. Driven by her intellectual curiosity for data and emergent information technologies, and her passion for designing and building things, she has oriented her career in the direction that integrates research, teaching, and practice in the realm of information science. Her research focuses on data and knowledge management areas such as data mining, data warehousing, and semantic technologies with an emphasis on exploring the synergies between information systems and data analytics. She was a data scientist in the industry with hands-on experience in data analytics, data mining, and big data platforms.

Manoj Abraham Thomas is an Associate Professor in the Department of Information Systems at Virginia Commonwealth University. He conducts research in diverse settings involving uncertain environments, non-traditional users, and unconventional application of technological solutions. He has been involved in ICT research in Brazil, Botswana, India, Haiti, Portugal, and the United States. His research has been published and presented internationally. He rides and works on motorcycles when he wants to get away from the digital realm.

Kweku-Muata Osei-Bryson is Professor of Information Systems at Virginia Commonwealth University. His research areas include: Data Mining, Decision Support Systems, Knowledge Management, IS Security, e-Commerce, IT for Development, Database Management, IS Outsourcing, Multi-Criteria Decision Making. He has published in various leading journals including: Decision Support Systems, Information Systems Journal, Expert Systems with Applications, European Journal of Information Systems, Information Systems Frontiers, Knowledge Management Research & Practice, Information Sciences, Information & Management, Journal of the Association for Information Systems, Journal of Information Technology for Development, Journal of Database Management, Computers & Operations Research, Journal of the Operational Research Society, & the European Journal of Operational Research. He serves as an Associate Editor of the INFORMS Journal on Computing, as a member of the Editorial Boards of the Computers & Operations Research Journal & the Journal of Information Technology for Development, and as a member of the International Advisory Board of the Journal of the Operational Research Society.

Please cite this article as: Y. Li, et al., A snail shell process model for knowledge discovery via data analytics, Decision Support Systems (2016), http://dx.doi.org/10.1016/j.dss.2016.07.003