Developments in Knowledge Discovery Processes and Methodologies: Anything New?

Completed Research

Jeroen Baijens
Open University, the Netherlands
jeroen.baijens@ou.nl

Remko W. Helms
Open University, the Netherlands
remko.helms@ou.nl

Abstract

The process of turning data into knowledge is referred to as "knowledge discovery" (KD) and originated in the 1990s. Since that time many different process models and methodologies have been developed. A genealogy presented in 2010, showed how the different models evolved and presented a refined process model, which represents a synthesis of the models presented before. However, the rise of data analytics and big data have changed how organizations do business. The key to these changes is to use data and turn it into knowledge to create value for the organization. Therefore, this study aims to update our understanding of knowledge discovery processes by reviewing the research into KD processes since 2010 in order to understand if there have been considerable changes and developments in this field. The developments in KD process models and methodologies that were found are threefold: tasks, steps and agile practices.

Keywords

Knowledge discovery process, Process model, Process methodology, Agile practice, Big data.

Introduction

Nowadays organizations are interested in creating value from data by drawing on analytical techniques to convert raw data into actionable knowledge. This knowledge supports managerial decision-making and allows the organization to take actions that might help creating or sustaining competitive advantage (Provost and Fawcett 2013). The process of using data to create knowledge has already been studied in the 1990s and was referred to as "knowledge discovery" (KD), or "knowledge discovery in databases" by Fayyad, Piatetsky-Shapiro, and Smyth (1996). Today, practitioners and academics often use the term "data analytics" or "data science" interchangeably with the older term knowledge discovery (Chen et al. 2012).

The research program into KD which has started in late 1990 has resulted in an abundance of proposed process models and methodologies developed by academics as well as practitioners. The most well-known model is CRISP-DM and is developed by a consortium consisting of industry and academic representatives (Chapman et al. 2000). Mariscal, Marbán and Fernández (2010) reviewed the existing literature on KD process models and proposed a refined KD process model based on a synthesis of the existing process models and methodologies. The resulting model consists of 3 main processes and 17 sub-processes and is the greatest common divisor of the models they analyzed. Rather than a process model it is better called a framework since it only identifies the main and sub-process without further detailing them or providing a complete methodology. However, despite the abundance of models, a survey among data science professionals reveals that 82% of them did not use any existing process model and methodology for knowledge discover (Saltz, Wild, Hotz and Stirling, 2018). Critics of the process models and methodologies argue they are too rigid and do not support the iterative and open nature of most KD projects (Saltz, 2015).

This modest uptake might be caused by the fact that most models and methodologies are still very rudimentary and not fit every situation. Mariscal et al. (2010) called for further research into the KD process by further extending the models and methodologies by borrowing from other fields (e.g. software development). Since 2010, several studies have been conducted to further develop and extend the KD

process models and methodology. Many of these studies were inspired by the rise of big data and data analytics and aimed at developing or applying KD process models and frameworks across different industries and various types of data (Ahangama and Poo 2014; Li et al. 2016).

This study aims to provide an overview on the evolution of KD process models since the review by Mariscal et al. (2010). To this end, a systematic literature review of KD process models and methodologies was conducted in which we categorized the content of the articles using thematic analysis. In this way, we were able to gain a thematic overview with regards to current research in KD. As such, this paper will show how KD models and methodologies have evolved in current years. By focusing on the field of KD in its entirety, our efforts are complementary to prior reviews which have focused on illuminating specific areas of KD such as KD processes models for big data (Saltz and Shamshurin, 2016). Furthermore, as data-driven sustainable development is on the agenda for the digital society to pave the way towards digital transformation and sustainable societies (Pappas et al. 2018). This research contributes by presenting an overview that supports research on the development of sustainable data-driven processes.

The remainder of this paper is structured as follows. In section 2, the method of our review is described. Thereafter, we present our results. Finally, we have a discussion and conclusion which includes suggestions for future research.

Method

The literature review was conducted according to the guidelines presented in Okoli and Schabram (2010) and Webster and Watson (2002). Following these guidelines is essential to create a scientifically rigorous systematic literature review. Steps in this guideline include; purpose of the literature review, protocol and training, searching for literature, practical screening, quality appraisal, data extraction, analysis of the findings and writing the literature review. Therefore, this systematic literature review followed a process consisting of the following phases: search, selection, analysis, and synthesis. In this review the focus was on identifying articles which investigate the process and or methodology of knowledge discovery after 2010.

Search and Selection Process

The systematic review included peer-reviewed research articles published in academic outlets, such as journal articles and conference proceedings, within the Web of science, AIS eLibrary and IEEE Xplore database. To do so we formulated a search query with keywords and searched for the occurrence of these keywords within the title, abstract, and keyword sections of the articles. Due to different ways of how these databases work we used Web of Science to search on "Topic", the AIS electronic library to search on "Title", "Abstract", and "Subject" and the IEEE Xplore digital library to search on the "Title", "Abstract" and "Index terms". The query used to execute the search process is a combination of two sets of keywords with the first term being "knowledge discovery" and the second "process model". For the former search term we also included the following synonyms in our search: "data analytics" and "data science". While for the latter search term the following synonyms were included: "process view", "process methodology", "analytic process", "knowledge discovery process", and "data science process". We reviewed literature from 2010 onwards, to cover the literature after Mariscal et al. (2010) presented their refined knowledge discovery process. Furthermore, the year 2010 was also chosen as a cutoff point as it represents the time period when research into big data and data analytics was starting to accumulate. The search was conducted from September 3, 2018 to October 15, 2018 and resulted in a total of 595 unique articles (after removing duplicates).

We subsequently screened the title and abstract of the articles to determine their relevance to the systematic review. At this stage, studies were excluded if it was clear that they did not address the knowledge discovery process. The number of articles after the screening was 93. Each of these articles was subsequently fully assessed by one of the authors. During this screening, we included studies in our corpus if they either contained detailed information about the performance of the KD process or if they contained information on how to support the KD process. In addition, articles that were focused on processing data or studies that discussed technical aspects of data analytics like algorithms were excluded. Also, articles that referred to the process of implementing analytics were excluded as this process is not relevant to managing an analytics project. Furthermore, articles written in non-English languages and articles stemming from non-peer reviewed conferences or journals were excluded to ensure the quality of the papers. After this we had a set

of 30 articles that discussed the process of KD. Although some articles discuss a process model, they did not add anything new to existing process models. They mainly tested a process model in a specific context. Therefore, we excluded them and only included the articles that discuss adjustments to the existing process models. This led to a set of 6 articles. Based on this initial set of 6 articles we engaged in backward and forwards snowballing in order to identify articles that were not captured by our initial search. This resulted in 3 additional articles that were added to our set of articles. Our final set of articles consisted of 9 studies. An overview of the search and selection process is shown in figure 1.

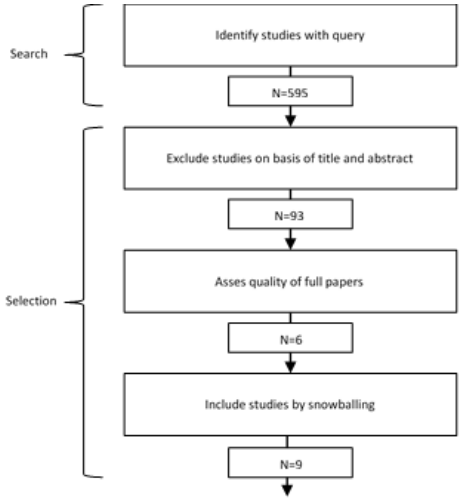

Figure 1 Search and selection process

Analysis and Synthesis of the Literature

The analysis focused on identifying the type of adjustments made to knowledge discovery process models and methodologies since the review presented by Mariscal et al. (2010). We used a concept-centric approach, or thematic analysis, to synthesize the literature. This helped with grouping the key findings from the literature study (Webster and Watson 2002). We identified adjustments in three separate dimensions related to the KD process, namely: 'tasks', 'steps', and 'agile practices'. First, tasks contain certain activities that need to be done and have no specific sequence within certain step. Second, steps also named as phases, cover a set of tasks that need to be done and can have a specific sequence to follow. Last, agile practices in KD provide a certain way to approach a task or activity. Agile practices are results from certain agile principles. These principles are basic generalizations and recognized as true. The agile practices are the applications of this principle in a certain setting (Williams 2010).

Results

The developments that traditional KD process models and methodologies underwent vary from proposing new tasks, steps, or adding agile practices. Six process models and two methodologies concerning adjustments to KD process models were identified in the articles. The process models describe what needs to be done and adjust specific steps or tasks to fit a specific situation or context. On the other hand, two methodologies address how this should be done and present approaches to conduct KD.

Steps and Tasks

Four process models in the literature (i.e. Ahangama and Poo, 2014, 2015; Grady, 2016; Li et al., 2016; Angee, 2018) proposed adjustment in steps and tasks for KD process models. To facilitate the comparison with previous models we will contrast the adjusted models with the original model as proposed by Mariscal et al. (2010). This comparison is presented in table 1 in which the steps of the refined model by Mariscal et al. (2010) are presented in the most left column. The second column shows the steps of the CRISP-DM model and the other four columns show the four new models that propose adjustments to steps and tasks. Highlighted cells indicate the differences they propose in reference to Mariscal et al. (2010) refined model

these can be new step or tasks in these steps. The process models are mapped together on similar steps they have compared with the refined model. When there are similar tasks between models discussed they are matched and will be approach as similar steps. In the remainder of this section the adjustments in steps and tasks are discussed in more detail.

First *life cycle selection* is put as a separate step by Mariscal et al. (2010). The process model proposed by Li et al. (2016) does not provide a step for this, but includes this as a task to determine which project management methodology should be used in the *business understanding* step. They add this task because the iterative nature of KD may require a more agile or hybrid methodology rather than the more traditional waterfall approach. Consequently, organizations should choose the best project management methodology which fits best with their culture and project type.

<table>
<tr><th>Mariscal et al. (2010)</th><th>CRISP-DM (Chapman et al., 2000)</th><th>(Ahangama and Poo, 2015)</th><th>(Li et al., 2016)</th><th>(Angee, 2018)</th><th>(Grady, 2016)</th></tr>
<tr><td>(1) Life Cycle Selection</td><td></td><td></td><td>Business understanding</td><td></td><td></td></tr>
<tr><td rowspan="3">(2)Domain Knowledge Elicitation</td><td rowspan="4">Business understanding</td><td>Project initiation</td><td rowspan="3">Business understanding</td><td>Conduct readiness assessment</td><td rowspan="4">Plan</td></tr>
<tr><td rowspan="3">Domain understanding</td><td rowspan="3">Understand business</td></tr>
<tr><td>(3)Human Resource Identification</td></tr>
<tr><td>(4)Problem Specification</td><td>Problem formulation</td></tr>
<tr><td>(5)Data Prospecting</td><td>Data understanding</td><td>Data understanding</td><td>Data understanding</td><td>Understand data</td><td>Collect</td></tr>
<tr><td></td><td></td><td>Conceptualization</td><td></td><td></td><td></td></tr>
<tr><td>(6) Data Cleaning</td><td rowspan="3">Data preparation</td><td rowspan="3">Data Preparation</td><td rowspan="3">Data preparation</td><td rowspan="6">Build prototype</td><td rowspan="6">Curate</td></tr>
<tr><td>(7) Preprocessing</td></tr>
<tr><td>(8) Data Reduction and Projection</td></tr>
<tr><td>(9) Choosing the DM Task</td><td rowspan="4">Modeling</td><td rowspan="4">Data Modelling</td><td rowspan="4">Modeling</td></tr>
<tr><td>(10) Choosing the DM Algorithm</td></tr>
<tr><td>(11) Build Model</td></tr>
<tr><td>(12) Improve model</td></tr>
<tr><td>(13) Evaluation</td><td rowspan="2">Evaluation</td><td rowspan="2">Validation</td><td rowspan="2">Evaluation</td><td rowspan="2">Evaluate prototype</td><td rowspan="3">Act</td></tr>
<tr><td>(14) Interpretation</td></tr>
<tr><td>(15) Deployment</td><td>Deployment</td><td>Presentation</td><td>Deployment</td><td></td></tr>
<tr><td>(16) Automate</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>(17) Establish On-going Support</td><td></td><td>Presentation</td><td>Maintenance</td><td></td><td>Act</td></tr>
</table>

Table 1 Overview of process model steps

The *domain knowledge elicitation* step should include a task for cost-benefit analysis which assesses if the expenses associated with the KD process are justified by the estimated value that will be created (Grady 2016). Furthermore a task for assessment of the analytics maturity and a task for enterprise knowledge acquisition by thorough reviewing enterprise content management systems and having conversations with business users and analysts. This helps combining explicit and tacit knowledge on problem domain (Li et al. 2016). Furthermore, Ahangama and Poo (2015) propose that a task is needed that determines compliance needs. Therefore, it is essential to have policies, procedures and guidelines in place. This task is necessary to stay compliant with national and international privacy laws and mitigate litigation risks (Grady 2016).

In the *human resource identification* step, Ahangama and Poo (2015) propose to add a task that determines stakeholder requirements which identify the stakeholder and their role in the project. In the step *problem specification,* Li et al. (2016) suggest adding a task which formulates the business problem that can be solved with a what, why, and how questions. Further, they suggest adding a problem decomposition task, which divides the main problem into components and determines the project boundaries (Grady 2016; Li et al. 2016). In the *data prospecting* step a task to have an understanding of the big data assets with respect to volume, velocity and veracity is crucial. This ensures that the organization can examine the challenges from the business and modeling requirements (Angee 2018; Grady 2016; Li et al. 2016). Moreover, dynamic access to a dataset is recommended to make it easier to ingest a dataset, and that not only the data provider has access to it (Grady 2016). The *data cleaning* step has tasks that should include an impact analysis which indicates the degree to which the quality of the data will affect the results of the analytic activities (Grady 2016). Further, a task that takes privacy into account when integrating and merging different data sets is proposed (Grady 2016). Moreover, ensuring alignment between data transformation and business requirements is a necessary task (Li et al. 2016). Where data is transformed based on the input from earlier steps to fit their requirements.

In the *build model* step a task should plan the development of a prototype model with the help of a workflow plan (Angee 2018). This helps with understanding the activities that need to be done in the team and provides insight into the required input and desired output. Furthermore, a repository of modeling rules for different modeling techniques is needed to help with decision support (Li et al. 2016). In the *improve model* step, when the modeling technique is applied the process model should have a task that assesses the analytics effort in order to decide whether a more efficient technical solution is possible (Grady 2016). For the *evaluation* step a task for the construction of a testing scenario is proposed for a model with unclear business objectives (Angee 2018; Li et al. 2016). In the *interpretation* step the task that is added is to ensure effective communications with all stakeholders in order to ensure achieving the business objectives (Li et al. 2016). Grady, (2016) suggests communicating these results by the creation of infographics or interactive dashboards. Finally, tasks for the e*stablish on-going support* step are related to the maintenance of the models. Four tasks are prosed for this step. First, establishing a maintenance process which describes the activities in the maintenance plan (Li et al. 2016). Second, a task to store all the details of the analytics activities like data input and transformation, modelling technique, performance measures, and business performance measures (Li et al. 2016). Third, deployed analytics models need procedures and guidelines on when and how changes in these models are implemented (Li et al. 2016). As environments change the models could use updates that take into account new factors (Ahangama and Poo 2015). Last, the model needs monitoring and should gather information to facilitate maintenance and future evolution. This can be achieved by gathering user feedback and keeping up-to-date with security regulations (Ahangama and Poo 2015; Grady 2016; Li et al. 2016). An overview of all the tasks that are added for the overall process is given in table 2.

Furthermore, there is one step identified that is not part of Mariscal et al. (2010) refined process model. The *conceptualization* step focuses on the exploration of variables that will be used and the relations between those variables (Ahangama and Poo 2015). The reason for this is that the analytic technique should not depend on the data available, but should have the organizations goal in mind. For this purpose it uses a scheme from the real world to create an abstract image of a specific area. With the help of a literature review and research questions, a conceptual model is created. This is not to be confused with the analytical model created in the modeling step, but theories in the represented area are selected used to describe the variables in the model. This will avoid using an abundance of different variables in the hope to find interesting relations. Tasks belonging to the *conceptualization* step are a literature review on the domain goal, formulate a research question and the development of a conceptual model with description of the variables used in the model. In addition, when dealing with statistical problems a hypothesis for each research question is needed (Ahangama and Poo 2015).

Agile Practices

Agile methodologies and practices are gaining more attention in KD projects (Saltz et al., 2018). The use of agile methodologies or practices is customary in software development. The main benefit of agile practices is that they enable organizations to better deal with volatile requirements which often result from operating in dynamic environments. By supporting fluid communication between stakeholders, agile techniques

allow organizations to quickly react to changing environments and help with generating returns on their analytics investments (do Nascimento and de Oliveira 2012).

Five process models in the literature (i.e. do Nascimento and de Oliveira, 2012; Ahangama and Poo, 2015; Li et al., 2016; Grady, Payne and Parker, 2017; Schmidt and Sun, 2018) proposed adding agile practices for KD process models. These practices can be proposed during specific steps or tasks. For instance, *pair programming* is a practice that can be used for the *modeling* step in a KD project, where two persons work together on creating an analytic model (Saltz and Shamshurin, 2017). One of the two is the "driver" and writes the code, and the other is the "observer" and reviews what is written and whether that is appropriate for the main goal. *Pair programming* helps programmers communicate efficiently, facilitate effective knowledge sharing, helps junior data scientists quickly learn from their senior colleagues, and helps with creating boundary conditions. The use of the *pair programming* practice is also proposed during the *data understanding* and *evaluation* step (Schmidt and Sun, 2018). In addition, *test-driven development* is an agile practice were an early code is written and tested to satisfy a set of user criteria (Williams 2010). Thus, a test case is first created and then the code is created to pass the test. This guarantees that the created code is constantly tested and leads to short development cycles. The use of the *test-driven development* practice is proposed during the *evaluation* step (Schmidt and Sun, 2018). Moreover, *continuous integration* is an agile practice where different developers integrate code from each other early in the process. It is put in one repository and tested to detect problems in the code. This integration is done on a regular basis. The use of *continuous integration* practice is proposed during the *data understanding* or *evaluation* step (Schmidt and Sun, 2018).

Mariscal et al. (2010)	**Tasks**
(1) Life Cycle Selection	-Determine project management methodology
(2)Domain Knowledge Elicitation	-Assessment of analytics maturity -Enterprise knowledge acquisition -Determine if expense justify the estimated value created -Policies, procedures and guidelines described for privacy
(3)Human Resource Identification	-Decide stakeholders requirements
(4)Problem Specification	-Formulate business problem -Determine organizational boundaries
(5)Data Prospecting	-Ensure understanding of the big data assets -Ensure dynamic access to a dataset
(6) Data Cleaning	-Determine how data quality effects analytics results -Take privacy into account with fusion of data -Alignment data transformation with business requirements
(11) Build Model	-Workflow of prototype -Describe modeling rules
(12) Improve model	-Assess analytics on whether a more efficient technical solution is possible
(13) Evaluation	-Construct testing scenario's
(14) Interpretation	-Effective communication with stakeholders
(17) Establish On-going Support	-Deploy changes in analytic models -Arrange maintenance life-cycle -Model monitoring across process -Store analytics results

Table 2 Proposed tasks for knowledge discovery process

Some of the agile practices proposed in KD process can be used across all steps. These are; *time-boxed iteration, user story, standup meetings* and *sprint efforts*. First, *time-boxed iteration* can be used as an agile practice to help the team with providing predictable incremental value to their stakeholder. It sets a timeframe to deliver incremental value within a specific period (Li et al. 2016; do Nascimento and de Oliveira 2012). Second, a *user story* is a short description of what the end user desires from the end product. The end user is often not part of the KD team. Thus, the *user story* ensures that they can influence the development of the end product (Schmidt and Sun 2018). Third, in *standup meetings* the KD team comes together for a short timeframe to discuss the work efforts for that day. The discussion is focused on what is done, what is needed and what the barriers are. Schmidt and Sun (2018) propose this practice during the *data understanding step*. However, Li et al. (2016) propose that the use of daily *standup meetings* is beneficial during the whole KD process. Last, in a *sprint* practice there is a certain time frame for the work to be done. When this is finished it is reviewed and presented to the stakeholders. The review and feedback

from stakeholders is then used as input for another *sprint* (Grady et al. 2017; Larson and Chang 2016; Schmidt and Sun 2018).

KD teams can combine a set of different agile practices to create their own hybrid agile methodology. However, there are also predefined agile methodologies that combine a set of agile practices (Williams 2010). The use of agile methodologies for knowledge discovery without specific process steps is proposed by Saltz, Heckman and Shamshurin (2017). They experiment with the *Kanban* and *Scrum* methodology in data science teams. First, in *Scrum* the overall project is divided into a set of smaller projects. Each smaller project is carried out in a sprint of two weeks. During the execution of this sprint, the team is at that moment not allowed to implement suggestions for improvements on the planned work. The suggestions that arise during project execution are saved for the next sprint. Next, the *Kanban* methodology makes use of a "Kanban board" which shows the work to do. All tasks that belong to a phase are put on the board. With this the team can create a prioritization list of tasks. The board highlights tasks that can be done simultaneous and leads to less problems with bottlenecks during the process (Saltz, Heckman and Shamshurin, 2017). An overview of the agile practices per step of CRISP-DM is given in table 3.

Step	Agile practices
1) Business understanding	User story, Time-boxed iterations, Sprint efforts, Stand up meetings
2) Data understanding	Continuous integration, Pair programming, User story, Time-boxed iterations, Sprint efforts, stand up meetings
3) Data preparation	User story, Time-boxed iterations, Sprint efforts, Stand up meetings
4) Modeling	Pair programming, User story, Time-boxed iterations, Sprint efforts, Stand up meetings
5) Evaluation phase	Continuous integration, Test driven development, Pair programming, User story, Time-boxed iterations, Sprint efforts, Stand up meetings
6) Deployment	User story, Time-boxed iterations, Sprint efforts, Stand up meetings

Table 3 Agile practices

Furthermore, iteration is a crucial element in using agile in KD process. The life cycle decides the sequence on which tasks need to be done. Process models often have a waterfall life cycle. Therefore, feedback loops provide a way to iterate the process and to create an improved output (Marbán et al. 2009). Different authors process new feedback loops to provide more options for iteration at different steps. While the traditional CRISP-DM only provide feedback loops after the *data understanding, modeling* and *evaluation* step, Angee (2018) proposed feedback loops from different steps toward the *business understanding* step. In addition, Ahangama and Poo (2014, 2015) divide two main cycles of iteration. One between the *domain understanding*, *data understanding* and *conceptualization* and the other between *data preparation*, *modeling* and *evaluation*. Furthermore, loops across all steps are proposed by Schmidt and Sun, (2018) and Li et al. (2016), in order to promote iteration.

Discussion and Conclusion

Since the study by Mariscal et al. (2010), there have been several researches that propose new or additional tasks, steps or agile practices. The adjustments proposed are often compared to the traditional CRISP-DM (Angee 2018; Grady 2016; Li et al. 2016). An essential driver for proposing the adjustments is to make the CRISP-DM model useful in a big data analytics context. The results from the literature review identified that big data is a common theme to adjust a process model or methodology. Big data provides organizations opportunities, but also provokes many challenges to the KD process, as it makes the process complex to follow (Angee 2018; Grady et al. 2017; Li et al. 2016). Big data leads to large volume, high velocity and variant sources of data. The large volume causes more technical challenges to use data instead of traditional volume (Laney 2001). Furthermore, the high velocity, assures for a need in faster knowledge creation delivery and the high volume leads to increased responsibility in governance (Li et al. 2016).

However, steps that are proposed as improvements to the existing process models (mainly CRISP-DM) cause some unclearness on their added value. A closer inspection of the activities in identified steps reveal that these activities are also part of the CRISP-DM model, but they are not considered a separate step in CRISP-DM. For example, the CRISP-DM model does not have a distinct problem formulation step, but there is an activity in the *business understanding* step that addresses the business problem by formulation business objectives. Thus, adding a separate step especially for *problem formulation* seems unnecessary. However, the dynamic environment in a big data context requires a distinct problem formulation due to the

complexity of this context. This complexity is caused by high volume, velocity and variety of sources which increase the technical challenges, the faster need of knowledge and increased responsibility. Therefore, Li et al. (2016) split the business understanding from the CRISP-DM model in a distinct problem formulation step similar to the problem specification of Mariscal et al. (2010). However, their model does not give detailed information on how to handle this step. In the *problem formulation* step the goal is to formulate a business problem that needs to be solved with a knowledge discovery project. A well-formulated problem statement will contribute to a clear focus and ultimately helps to solve the business problem (Li et al. 2016). The current CRISP-DM methodology lacks such a delineated problem formulation step. The step ensures that an organization has a clear idea on what, why, and how they approach their knowledge discovery activities. The problem itself can be identified by the organizational requirements or new identified ways of doing data analytics that could be worthwhile (Ahangama and Poo 2014, 2015). A well-formulated business problem will help in managing the expectation form top management.

Similarly, the use of big data increases the complexity of model deployment and maintenance. This is due to the increase in technical challenges, faster need and increased responsibility it brings. Mariscal et al. (2010) already provided a *establish on-going support* step to take care of this. However, they do not provide details on the specific activities in this step. Li et al. (2016) propose to distinguish this step form the *deployment* step as a separate *maintenance* step. The *deployment* step is often perceived as endpoint in the process, where implementing change is difficult. Thus splitting this step could contribute to implementing changes easier. The tasks proposed in this step are very similar to the already existing tasks in CRISP-DM. The proposed task, which is to guide business users on how to deploy changes to the models is covered in CRISP-DM by the 'plan monitoring and maintenance' task. This task determines when and what should happen when the model results should not be used anymore. (Chapman et al. 2000). Therefore, adding this step is not a contribution to a new process model. However, it did not have a task to established a maintenance process. Thus, a formalized one is valuable, where the CRISP-DM model only mentions that a maintenance plan is needed (Li et al. 2016). Furthermore, CRISP-DM includes monitoring of the models to assess their performance. However, monitoring on security and feedback is not included in the CRISP-DM model (Ahangama and Poo 2015; Grady 2016; Li et al. 2016).

Furthermore, some tasks proposed seem to be already included in the traditional CRISP-DM models. The tasks 'Determine if expense justify the estimated value created', 'Assess analytics effort' and 'data quality', proposed by (Grady 2016) are already mentioned in CRISP-DM model as; 'costs and benefits analysis', 'assess model', and 'Verify Data Quality' tasks (Chapman et al. 2000). Thus, the value that these tasks add to the process model is not clear as they appear to be very similar to existing tasks. Also, the communication with stakeholders is mentioned in CRISP-DM during the final presentation in the produce final report task. However, CRISP-DM does not explicitly discuss to communicate these results through visualization via infographics or interactive dashboards (Grady 2016).

Another theme that is retrieved from the literature on process models is to adjust process models for the specific healthcare environment. One process model was designed to deal with the diversity in the healthcare ecosystem and the diversity of available health analytic techniques (Ahangama and Poo 2014, 2015). The dynamic context and patient-centric field cause that requirements variate rapidly. This results in existing approaches that do not seamlessly work for health analytic projects. Therefore, Ahangama and Poo (2015) propose a *conceptualizations* step in their model which is not present in the refined model of Mariscal et al. (2010) and similar activities are not addressed in CRISP-DM. Although this model can be generalized the step seems most relevant in a health care setting instead of a business environment. In a healthcare setting taken certain decisions can impact the quality of care for a patient. These impacts can be minimized by a proper conceptualization of the problem with the help of theory. This is not the case in a business setting, where theory is less critical and often not available. Furthermore, in the *human resource identification* step, Ahangama and Poo (2015) discusses the importance of a task to determine stakeholders requirements. The identification of stakeholder as human resources in a project is discussed in CRISP-DM, but not how these stakeholders are related to the project.

Various proposed tasks are new to the existing literature in process models and seem to have added value. Also, several tasks that are similar to existing ones, but expanded are perceived as valuable. However, some proposed tasks are already existing and unnecessary to propose as new tasks, as they do not cover new elements. An overview of all tasks that are expanded and tasks that were already available in CRISP-DM is given in table 4.

Tasks	Existing/ expanded
-Determine if expense justify the estimated value created	Existing in CRISP-DM
-Decide stakeholders requirements	Expanded
-Determine how data quality effects analytics results	Existing in CRISP-DM
-Assess analytics on whether a more efficient technical solution is possible	Existing in CRISP-DM
-Effective communication with stakeholders	Expanded
-Deploy changes in analytic models	Existing in CRISP-DM
-Arrange maintenance life-cycle	Expanded
-Model monitoring across process	Expanded

Table 4 Existing and added tasks

Agile practices like *time-boxed iteration, user story, standup meetings* and *sprint efforts* can improve the efficiency across the whole KD process, and the practices like *pair programming, test driven development, and continuous integration* are suggested as helpful during certain steps. Also adding more feedback for more option for iteration within the process will contribute to a more effective KD process model. While these practices are compared to the steps where they are performed, it is still unclear for which specific tasks they could be used an how they are evaluated. Further research is needed in how these practices can improve the performance of these tasks.

This paper contributes in presenting on an overview of the suggested improvements to KD process models and methodologies. It helps in choosing the steps to take, the tasks to do and which agile practice to add during the KD process. Furthermore, it gives academics guidance in evaluating the adjustments proposed in KD process models in order to continue research in developing these models. Moreover, governance on analytics activities is now an interesting new area of research which needs attention (Espinosa and Armour 2016). This paper contributes in giving an overview of tasks that should be considered in creating a governance structure. Still there are several limitations in this literature review. We only found a limited amount of papers on process models that proposed adjustments on task, steps, or agile practices. Therefore, the generalization of the findings is difficult. Furthermore, we did not have a look into practitioner literature, which could provide different result in developments that are already in use. Drawing from the literature discussed we identified that future research should focus on application, validation, evaluation and testing the process methodologies in different industries, data types or agile practices. This needs to be done on a larger scale with a more significant sample to test it statistically or by experiment (do Nascimento and de Oliveira, 2012; Ahangama and Poo, 2015; Li et al., 2016; Saltz, Shamshurin and Crowston, 2017; Angee, 2018; Schmidt and Sun, 2018).

Acknowledgements

This research was supported by the Province of Limburg, The Netherlands, under grant number SAS-2014-02207.

References

Ahangama, S., and Poo, D. C. C. 2014. "Unified Structured Process for Health Analytics," *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering* (8:11), pp. 768–776.

Ahangama, S., and Poo, D. C. C. 2015. "Designing a Process Model for Health Analytic Projects," in *PACIS 2015 Proceedings*. 3.

Angee, S. 2018. "Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-Organization Big Data & Analytics Projects," in *International Conference on Knowledge Management in Organizations* (Vol. 877), pp. 613–624.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. 2000. "Crisp-Dm 1.0," *CRISP-DM Consortium*. (https://doi.org/10.1109/ICETET.2008.239).

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165–1188.

Espinosa, J. A., and Armour, F. 2016. "The Big Data Analytics Gold Rush: A Research Framework for

Coordination and Governance," in *Proceedings of the Annual Hawaii International Conference on System Sciences* (Vol. 2016–March), pp. 1112–1121.
Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "Knowledge Discovery and Data Mining: Towards a Unifying Framework.," *Int Conf on Knowledge Discovery and Data Mining*, pp. 82–88.
Grady, N. W. 2016. "Knowledge Discovery in Data Science," in *2016 IEEE International Conference on Big Data*, pp. 1603–1608.
Grady, N. W., Payne, J. A., and Parker, H. 2017. "Agile Big Data Analytics: AnalyticsOps for Data Science," in *Proceedings 2017 IEEE International Conference on Big Data*, pp. 2331–2339.
Laney, D. 2001. "3D Data Management: Controlling Data Volume, Velocity, and Variety," *Application Delivery Strategies Mete Group*. (https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf).
Larson, D., and Chang, V. 2016. "A Review and Future Direction of Agile, Business Intelligence, Analytics and Data Science," *International Journal of Information Management* (36:5), Elsevier Ltd, pp. 700–710.
Li, Y., Thomas, M. A., and Osei-Bryson, K.-M. 2016. "A Snail Shell Process Model for Knowledge Discovery via Data Analytics," *Decision Support Systems* (91), Elsevier B.V., pp. 1–12.
Marbán, O., Segovia, J., Menasalvas, E., and Fernández-Baizán, C. 2009. "Toward Data Mining Engineering: A Software Engineering Approach," *Information Systems* (34:1), pp. 87–107.
Mariscal, G., Marbán, Ó., and Fernández, C. 2010. "A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies," *The Knowledge Engineering Review* (25:2), pp. 137–166.
do Nascimento, G. S., and de Oliveira, A. A. 2012. "An Agile Knowledge Discovery in Databases Software Process," in *The Second International Conference on Advances in Information Mining and Management Compliance*, pp. 343–351.
Okoli, C., and Schabram, K. 2010. "Working Papers on Information Systems A Guide to Conducting a Systematic Literature Review of Information Systems Research," *Working Papers on Information Systems* (10:26), pp. 1–51.
Pappas, I. O., Mikalef, P., Giannakos, M. N., Krogstie, J., and Lekakos, G. 2018. "Big Data and Business Analytics Ecosystems: Paving the Way towards Digital Transformation and Sustainable Societies," *Information Systems and E-Business Management* (16:3), Springer Berlin Heidelberg, pp. 479–491.
Provost, F., and Fawcett, T. 2013. "Data Science and Its Relationship to Big Data and Data-Driven Decision Making," *Big Data* (1:1), pp. 51–59.
Saltz, J., Heckman, R., and Shamshurin, I. 2017. "Exploring How Different Project Management Methodologies Impact Data Science Students," in *Twenty-Fifth European Conference on Information Systems (ECIS), Guimarães, Portugal, 2017*, pp. 2939–2948.
Saltz, J. S. 2015. "The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness," *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 2066–2071.
Saltz, J. S., and Shamshurin, I. 2016. "Big Data Team Process Methodologies: A Literature Review and the Identification of Key Factors for a Project's Success," *Proceedings - 2016 IEEE International Conference on Big Data*, pp. 2872–2879.
Saltz, J. S., and Shamshurin, I. 2017. "Does Pair Programming Work in a Data Science Context ? An Initial Case Study," in *2017 IEEE International Conference on Big Data*, pp. 2348–2354.
Saltz, J. S., Shamshurin, I., and Crowston, K. 2017. "Comparing Data Science Project Management Methodologies via a Controlled Experiment," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, pp. 1013–1022.
Saltz, J. S., Wild, D., Hotz, N., and Stirling, K. 2018. "Exploring Project Management Methodologies Used Within Data Science Teams," in *Twenty-Fourth Americas Conference on Information Systems, New Orleans, 2018*, pp. 1–5.
Schmidt, C., and Sun, W. N. 2018. "Synthesizing Agile and Knowledge Discovery: Case Study Results," *Journal of Computer Information Systems* (58:2), Taylor & Francis, pp. 142–150.
Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future : Writing a Literature Review," *MIS Quarterly* (26:2), pp. 13–23.
Williams, L. 2010. "Agile Software Development Methodologies and Practices," *Advances in Computers* (80), Elsevier Inc., pp. 1–44.