Applying Scrum in Data Science Projects

Applying Scrum in Data Science Projects

Jeroen Baijens
Department of Information Science
Open University
Heerlen, The Netherlands
jeroen.baijens@ou.n

Remko Helms
Department of Information Science
Open University
Heerlen, The Netherlands
remko.helms@ou.nl

Deniz Iren
Department of Information Science
Open University
Heerlen, The Netherlands
deniz.iren@ou.nl

***Abstract*—The rise of big data has led to an increase in data science projects conducted by organizations. Such projects aim to create valuable insights by improving decision making or enhancing an organization's service offering through data-driven services. However, the majority of data science projects still fail to deliver the expected value. To increase the success rate of projects, the use of process models or methodologies is recommended in the literature. Nevertheless, organizations are hardly using them because they are considered too rigid and they do not support the typical iterative and open nature of data science projects. To overcome this problem, this research suggests applying Agile methodologies to data science projects. Agile methodologies were originally developed in the software engineering domain and are characterised by their iterative approach towards software development. In this study, we selected the Scrum approach and integrated it into the CRISP-DM methodology for data science projects using a Design Science Research approach. This new methodology was then evaluated in three different case organizations using expert interviews. Analysis of the expert interviews resulted in a further refinement of the Agile data science methodology proposed by this research.**

Keywords—Data Science, Agile, Scrum

I. Introduction

Many organizations nowadays conduct data science projects to create valuable insights to improve decision making or enhance service offerings through the creation of smart services [1]. However, 85% of the projects that are executed fail to deliver the expected value [2]. To guide these projects towards successful results, the use of a process model or a project methodology is recommended [3]. A well-defined, repeatable process model or methodology helps practitioners in managing the tasks involved in executing these projects. However, in practice, 82% of data science teams do not use an existing process model or methodology to guide their projects [4]. Critics of the process models and methodologies argue they are too rigid and do not support the iterative and open nature of most Knowledge Discovery (KD) projects [5]. Therefore, more and more organizations apply Agile methods in data science projects to improve their success rate [6–8]. One well known Agile method often applied in data science projects is Scrum. Scrum is characterized by time-boxed sprints to deliver incremental value and consists of different events, artefacts, and roles [9]. Previous studies argue that the use of (elements of) the Scrum method improves the success rate of data science projects [10–12]. In comparison with other agile methods Scrum is considered useful for organizations that aim for early results, as this method focuses on constant iteration to deliver quick incremental value.

Existing research about the use of agile methods on data science projects, applied a mixture of agile methods, and not the complete Scrum method. Hitherto, to the best authors' knowledge, so far, no study has reported the application of the complete Scrum method consisting of events, artefacts, and roles. Previous studies typically added certain Scrum practices to existing process models, and no detailed explanation is given.

However, Scrum was often perceived unclear and difficult to use [13]. In Scrum, the users have to estimate task duration upfront and this is challenging because they do may not know how long a certain task takes [14].

Therefore, this study focuses on designing a complete Scrum method for data science projects (Scrum-DS). Scrum-DS uses elements of Scrum and applies them to the steps of CRISP-DM and evaluates this by demonstrating it to members of data science teams. This will provide a more detailed insight on which specific elements of Scrum contribute to improving the success rate of data science projects.

Hence, the research questions are as the following. "*How can the Scrum method be applied to improve the execution of data science projects in organizations?"* More specifically*, "how can Scrum events, artefacts, and roles be effectively used in data science projects?"*

A Design Science Research (DSR) approach was used to develop a tailored version of Scrum method for data science projects, i.e. Scrum-DS. The method is evaluated in terms of compatibility with data science projects in three different cases by expert interviews [16, 17].

The remainder of this paper is structured as follows. Section 2 presents the research background on Scrum and data science process models. Next, section 3 presents the related work on the field of agile in data science. Then, section 4 describes the DSR methodology of our study. Thereafter, section 5 presents the design of Scrum-DS, and section 6 provide details on the demonstration and evaluation. In section 7, a refined design of the artefact is presented. Finally, a conclusion is presented in section 8, including implications to science and industry and suggestions for future research.

II. Research background

In this section, we provide background on three Scrum elements; artefacts, events, and roles, as shown in Table 1. Furthermore, this section provides background on the most used data science process models (KDD and CRISP-DM).

A. Artefacts

The Scrum method consists of four different artefacts: user story, product backlog, sprint backlog, and increment.

First, the user story is a short description of a desire from the viewpoint of the end-user. As in traditional software development, a user story can be described as a feature of a software product [17]. In the end, user stories help to deliver fully realized work items in each iteration [18]. Therefore, the user story should be independent, valuable, estimable, testable, and realizable [19].

2378-1971/20/$31.00 ©2020 IEEE
DOI 10.1109/CBI49978.2020.00011

Second, the product backlog is a complete list of desires from the stakeholders concerning the product. It provides an overview of what the team can work on in future sprints. The desires are described in user stories. The product backlog is filled by the product owner with user stories together with the development team [20].

Third, the sprint backlog is a list of items to be developed during a sprint. The sprint backlog is created during the refinement based on the items of the product backlog. On the sprint backlog, there are items on which the team will work during the next sprint [7]. A user story can be put in a sprint backlog if it is small enough to be finished within one sprint [11].

Last, an increment is the deliverable of a sprint and consists of several user stories that together result in a working or a semi-finished product [21]. For the stakeholders, the increments are an indicator of the progress that has been made [14, 23].

B. *Events*

In Scrum five events are used, these include sprints, daily stand-up, retrospective, review, and refinement.

First, a sprint is a fixed period (1-4 weeks) wherein activities are executed. Each sprint has an upfront formalized sprint goal [11]. The sprints in software development projects are often used in activities that require the team to design, develop or implement software. The duration of the sprint in traditional software development projects takes two to four weeks to deliver incremental value. [19].

Second, in a daily stand-up, the project team has a daily meeting from approximately 15 minutes to reflect on the delivered work from the past 24 hours and to plan the work for the next 24 hours [20]. This provides them with insights on the progress of the sprint [23].

Third, in the sprint review, there is a meeting where the results of the sprint are presented to the stakeholders. This meeting takes approximately four hours and the team shows the increment that is created during the sprint [9].

Fourth, the sprint retrospective is a meeting at the end of a sprint in which the Scrum team reflects on the work and collaboration of the past sprint. After this meeting, the team defines process improvements to implement in future sprints. This event will typically last for approximately three hours [10].

Last, the refinement happens at the beginning of a sprint where the team meets together to discuss and priorities the new user stories [12]. The user stories are then combined to create a product and sprint backlog [7].

C. *Roles*

Traditional Scrum roles include Scrum Master, Product Owner, and Development Team.

The Scrum Master is knowledgeable of the Scrum method and has different responsibilities. Firstly, he facilitates team members by organizing the sprint refinement and sprint retrospective meetings. Secondly, he is responsible for avoiding barriers during the process and provides the required resources for the team. Thirdly, he has also a supportive role towards the product owner, the development team and the business [23]. Fourthly, he is responsible that everyone understands Scrum [19]. Lastly, he is also responsible that no additional items are added during a sprint [24].

The product owner is the person who uses his business knowledge to prioritize the items on the product backlog. He is the representative of the business and responsible for optimizing the value of the work [12, 20].

The development team is responsible for creating working products. The team should be small enough to act quickly, but also large enough to get work done [21]. Therefore, team size is recommended between 3 to 9 members. A crucial aspect of this team is that it works cross-functional, is self-organizing and has no hierarchy.

D. *Data science proces models*

To effectively engage in data science to create social or economic value, organizations have to overcome challenges at different organizational levels [25]. To overcome these challenges, one stream of research focused on process models and methodologies, which provide guidelines for conducting data science activities. Research into the use of these models and methodologies started in the late 1990s with the Knowledge Discovery in Databases (KDD) model. The KDD model consisted of five steps: data selection, data pre-processing, data transformation, data mining, and data interpretation/evaluation [26]. Further research on this model has resulted in an abundance of proposed process models and methodologies [3].

The most well-known process model for data science is the CRISP-DM model and was developed by a consortium consisting of industry and academic representatives [27]. The model provides a set of six steps with tasks that need to be performed to deliver value [3].

First, the business understanding step ensures that from a business perspective there is a clear understanding of the objectives and requirements. Second, the data understanding step is to get familiar with the data, receive first insights and spot data quality problems [3, 27]. Third, the data preparation step covers all the tasks that are related to constructing the final data set that is used for modelling. Fourth, in the modelling step, the right modelling technique is chosen and applied on the data [3, 27]. Fifth, the evaluation step ensures that there is a detailed evaluation of the model that is built in the previous step. Therefore, there is a check whether the model meets the business objectives which were formulated in the business understanding step [27]. Last, in the deployment step, the created model is applied in the organization. This can be in the form of a report or a smart service [27]. Despite the detailed description, CRISP-DM is not an answer to all managerial and cultural barriers related to data science.

TABLE I. SCUM DATA SCIENCE ARTEFACTS, EVENTS AND ROLES

Artefacts	User story
	Product backlog
	Sprint backlog
	Increment
Events	Sprint
	Daily stand-up
	Sprint review
	Retrospective
	Sprint refinement
Roles	Scrum Master
	Product owner
	Development team

III. Related work

CRISP-DM is in practice often executed as a waterfall approach where a project is conducted by going through a sequence of steps. Although CRISP-DM was intended to be an iterative model, evidence suggests that it has been used in a rather waterfall-like approach [3, 28]. In more recent publications, improved versions of CRISP-DM have been proposed by adding steps or tasks (e.g. problem formulation, maintenance) [6, 8, 10, 29, 30]. These new process models were introduced to cope with the specific challenges in big data projects or in healthcare settings. Moreover, more iteration between steps has also been proposed in these new process models and methodologies [29]. In addition, to improve efficiency the use of Agile practices alongside a waterfall approach is recommended during a project. This development led to more hybrid methodologies combining both waterfall and Agile approaches.

The use of Agile approaches in data science projects gained popularity in recent years [8]. They facilitate volatile requirements and allows to quickly react to changing environments [10, 11]. This provided more flexibility during a project and improved the effectiveness within a project. Examples of these Agile approaches are Kanban and Scrum. The Kanban method 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 prioritized list of tasks. The board highlights tasks that can be executed simultaneously and leads to fewer bottlenecks during the process [13].

Previous studies applied different elements of Scrum method in data science projects. For example, in one study a method is created where all data science activities are executed in a set timeframe to deliver incremental value within a specific period [11, 12]. However, the effectiveness of his method was never measured.

Another study used KDD and CRISP-DM as waterfall process models and added elements of Scrum [10]. For example, they used user stories to ensure that the end-user can influence the development of the end product. They stated that "Listening to the users regarding how they planned to use the models and writing them down as stories helped data modellers understand and clarify the business requirements of the projects" [10]. Furthermore, they also made use of daily stand-up meetings and sprints.

IV. Research methodology

The research methodology chosen for this study is DSR as is it gives the possibility to apply and test an artefact in a real-life setting. Furthermore, DSR is an effective problem-solving methodology for the design of artefacts to make research contributions, using evaluation, communication, and scientific rigour practices [31].

Design science develops artefacts that are designed to interact in a problem context [31, 32]. The problem we aim to solve is that data science projects do not deliver their expected value. To achieve a solution for the problem we aim at creating an artefact by designing a methodology for using Scrum in a data science project.

The DSR methodology suggests the following steps for the development and evaluation of an artefact [33]:

1) Identify the problem and motivate

Concerning the research problem which is already discussed in section 1, organizations fail to deliver the expected value of a data science project. In addition, they struggle to use a process model or methodology to guide these projects. For them, it is unclear how such process models or methodologies could help them run these projects.

2) Define the objectives of a solution

The main objective of this study is to design a Scrum-based data science project method that organizations can use to guide their projects. In order to use a data science methodology, a recent study explored that the criteria 'compatibility' has an important influence why a data science methodology is used [15]. Compatibility of a project methodology means that the methodology should be feasible as otherwise, it has no purpose to exist (Feasibility), and it should be able to adjust the work in progress dynamically give speed and simplicity to development (Flexibility)[16]. This study will show how elements of Scrum are used in the CRISP-DM process model to create a method that satisfies these criteria.

3) Design and development

For this study we present our artefact design, i.e. Scrum-DS. The design is based on literature concerning the use of Agile in projects and literature on data science projects, which was collected by three students. After the literature review, the students designed their own Agile data science project methodology. Within their design, they all used different Scrum elements and integrated them in CRISP-DM.

In the next stage, they demonstrated their designs in different organizations by expert interviews. After the demonstration, the three designs were compared to each other and integrated by the lead researcher to the Scrum-DS method. Scrum-DS uses Scrum elements that were present in all three designs, i.e. artefacts, roles, and events.

4) Demonstration

For the demonstration of Scrum-DS in empirical setting, we conducted 14 expert interviews at three different organizations. The expert interviews were used to present the design of Scrum-DS. This was done by discussion how the Scrum events, artefacts and roles fit with the CRISP-DM steps. After the presentation of Scrum-DS, the participant reflected on the 'compatibility' criteria. This provided valuable input on how the participants perceived Scrum-DS. The interviews were conducted by three students that were connected to the organizations. At the start of the interviews, Scrum-DS was explained by the researcher. The participant was free to ask questions or make remarks on the method, which resulted in an open discussion of the method. During the interviews, the researchers were guided by an interview guide. The interview guide consisted of questions on Scrum elements applied to CRISP-DM. The presentation and reflection took approximately one hour per participant.

In two organizations, additional insights were collected using focus groups. These focus groups were conducted in the form of a workshop on the Scrum method bringing experts together who were previously interviewed. In the workshop, there was a discussion concerning the 'compatibility' of the Scrum method. Each session was hosted by one of the students

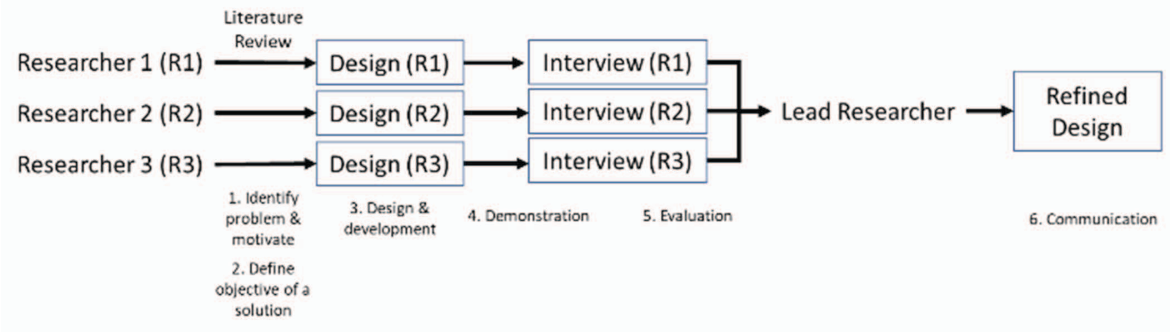

Fig. 1. DSR steps for the development of Scrum-DS

5) Evaluation

To evaluate the compatibility of Scrum-DS all collected data from the interviews and workshops was analysed. Therefore, the interviews and focus groups were recorded on a voice recorder and transcribed after demonstration. Analysing the interview data aimed at finding empirical evidence for Scrum-DS. More precisely, we looked for mentions in the interviews of artefacts, events, and roles of Scrum. To analyse the collected data, we went through a process of coding. For this purpose, we used a deductive approach, which allows using a theoretical framework for the analysis of qualitative data [38, 39].

The deductive approach involved the use of a priori codes to start the coding process and these codes were derived from the four artefacts, five events, and three roles. These 12 codes were used for one round of coding to mark portions of the interview data that relate to a specific Scrum element. In the end, the codes for each element of Scrum were summarized into more general observations. The lead researcher, who was not involved in the data collection, performed the coding.

By analysing the derived opinions on the artefact. We created summaries of the perception for all three types of Scrum elements. Based on the summaries we decided whether there was consensus on the compatibility criteria of the Scrum elements.

6) Communication

The last step involves communication of the findings from the artefact.

In this research, we follow this approach and an overview is presented in figure 1. The execution of the different steps is provided in the following sections.

V. Design and development

In this section, we will elaborate on how all Scrum elements are applied to CRISP-DM for the design of Scrum-DS, as shown in table 2.

A. Artefacts

The user stories in Scrum-DS can be described as an added feature to a data science product or service [17]. Moreover, a user story in a Scrum-DS can also be described as a sub-question. This sub-question is part of a bigger question that solves a business problem [20]. The creation of these user stories is done during the business and data understanding steps from CRISP-DM. This provides an overview of all activities required to deliver a data-driven solution.

The product backlog in Scrum-DS can be used similarly as in other fields, but items in a product backlog can deliver insights instead of a working product [7]. The creation of the product backlog is done during the business and data understanding steps from CRISP-DM.

To deliver incremental value for each sprint in Scrum-DS the sprint backlog should always include data preparation and modelling activities [20]. The creation of the sprint backlog is done during the business and data understanding steps from CRISP-DM.

The increment in Scrum-DS can be a data-driven product or can be some insights that help to solve a business problem [36].

B. Events

In Scrum-DS a sprint of 4 weeks is preferred [7, 19]. This longer time frame is required because data science projects are often dependent on the work of persons outside the team. For example, this happens when the development team does not have access to the right data during data preparation. In this situation, they first have to arrange access to the right data set [21, 35, 36].

The sprint is used in combination with the activities of the data preparation and modelling steps. For the reason that work is these steps can be cut in small parts to define user stories. This helps the team to divide the user stories and work individually on the creation of a product. Furthermore, in these steps, the organization already has a defined solution from the previous steps which is crucial to go into a sprint.

The daily stand-up in Scrum-DS is effective because the frequent communication can contribute to making the best decision for a solution together [18]. Especially in data science where there may be multiple options to tackle a problem. The daily stand-up is held during the data preparation and modelling steps. This makes the sprint events in these steps more effective.

For Scrum-DS, the refinement event will often result in new user stories with requests for additional data to make the model more accurate [21]. After the data understanding, it will take place before every sprint and adjusts the user stories in the product and sprint backlog.

During the sprint review in Scrum-DS, all the participants should be aware that an early data science model is not as accurate as required for the end product. The sprint review in Scrum-DS is held before the sprint retrospective during the evaluation step.

TABLE II. SCRUM-DS

CRISP-DM step	Business understanding	Data Understanding	Data preparation	Modelling	Evaluation
Events	• Refinement		• Sprint • Daily stand-up		• Sprint retrospective • Sprint review
Artefacts	• User stories • Product backlog • Sprint backlog		• Increment		
Roles	• Product Owner • Scrum Master • Development Team		• Scrum Master • Development Team		• Product Owner • Scrum Master • Development Team

After finishing the evaluation step, a new iteration of Scrum-DS is triggered. In the new iteration, the business and data understanding steps will refine the user stories, product and sprint backlog. This allows the development team to go into a new sprint of data preparation and modelling.

C. Roles

Three roles are applied in Scrum DS: Scrum Master, Product Owner, and Development Team. The Scrum Master is involved in every step of CRISP-DM and hosts the daily stand-up meetings.

The Product Owners responsibility is that the development team delivers a valuable product. Therefore, he manages the product backlog. Furthermore, he understands that the work in data science is creative and requires some trial and error [18]. He is involved in the evaluation and, business and data understanding steps.

For a data science project, the Development Team consists of the following roles; data miner, data modeller and data engineer [39]. The Development Team is involved in every step of CRISP-DM.

VI. DEMONSTRATION AND EVALUATION

In this section, the results of the demonstration and evaluation are discussed. The Scrum elements that are applied on CRISP-DM are evaluated on the compatibility criteria's (Flexibility and Feasibility), as shown in table 3.

A. Artefacts

According to the respondents, the user stories are an essential part of a Scrum-DS method. They provide the team with a clear description of the activities to work on. To create user stories, the work required in a data science project should be cut into pieces. To do that an estimation of the complexity is crucial because in data science you are building an algorithm and its complexity determines how long the development activity will take. The user stories should be created after the business and data understanding step because then the team has a clear view of what the end product will look like. With the creation of the user stories, all roles should be involved. Furthermore, the user stories should not be assigned to a specific team member but the product backlog. The team member who has time should take up a user story from the prioritized list from the product backlog.

Moreover, the addition of the product and sprint backlog is by the respondents also perceived valuable. With the product backlog, respondents mention that it is crucial to have it prioritized as soon as possible because it motivates the team to deliver. Based on this prioritized product backlog the development team should choose together with the Scrum Master the user stories that can be put in a sprint backlog. For the sprint backlog to succeed there must be an estimation on the amount of work per user story.

Furthermore, the respondents stated that it is challenging to deliver incremental value after a sprint due to its short period. The actual building of the model in a fixed period is not problematic, but the data preparation can be time-consuming. Therefore, with this method, there is no business value created in the first sprint. At the end of the sprint, there is no working product yet, and maybe only a finished data preparation.

Besides, data science projects can deliver a variety of increments. For example, a Business Intelligence solution in a dashboard or insight on a specific topic. However, it is challenging to decide whether an increment is finished. Is it finished when there is a complete dashboard, or is it finished when you collect data and calculated a percentage?

B. Events

Concerning the sprint event, there were different opinions on its usefulness in a data science project. For example, the sprint is useless when the project objectives are unclear. Therefore, it is important to have clear user stories defined after the business and data understanding step.

TABLE III. EVALUATION CRITERIA COMPATIBILITY OF SCRUM-DS

Scrum Elements		Compatibility			
		Consensus on Feasibility		Consensus on Flexibility	
		YES	NO	YES	NO
Artefacts	User story	X		X	
	Product backlog	X		X	
	Sprint backlog	X		X	
	Increment		X	X	
Events	Sprint		X	X	
	Daily stand-up	X		X	
	Retrospective	X		X	
	Sprint review	X		X	
	Sprint refinement	X		X	
Roles	Scrum Master	X		X	
	Product Owner	X		X	
	Development Team	X		X	

Furthermore, respondents question the possibility to use fixed periods to deliver an increment in a data science project. Despite that, some argued that depending on the problem it might be possible within two or three weeks if the data is available and infrastructure in place. The majority argued that a fixed period of 4 weeks is already challenging because the data preparation step is time-consuming. Sometimes, the data preparation step can take one whole sprint to get finished. As a result, no business value is delivered to the customer. The experts propose that it should be possible that a sprint can only consist of data preparation.

Furthermore, respondents identified issues that could arise during the use of sprints. For example, a small adjustment might be postponed to another sprint, as the team is not allowed to work on it. Thus, it can take three weeks when it is handled. Moreover, a disadvantage of the fixed sprint time is that when the work is finished and one week is left, the team is forced to only work on further improvements of the same user stories.

The daily stand-up was perceived as very useful by the respondents. It can help to identify early impediments that arise during the sprint. The daily stand-up is especially useful when people in the team have the same set of skill and the project itself is complex. Furthermore, as the daily stand-up provides an overall discussion there should be the possibility that you can discuss certain topics in-depth afterwards.

Concerning the refinement that happens before the start of the sprint, there were positive perceptions as well. It should be ensured that the user stories are created, and the sprint and product backlogs are filled. If the refinement happens for the second time after the first sprint than the user stories that were already formulated and the questions that pop-up during the last sprint needs to be handled.

Concerning the retrospective and the review, some experts argue that it is perhaps better to do it all in one meeting because then the stakeholders are also part of the meeting. However, the majority states it should be split because the retrospective focuses on the process during the sprint and the review is more on the content. Therefore, it makes more sense to follow these events upon each other.

C. Roles

According to the respondents, the Scrum Master is an essential role to use Scrum in data science. The Scrum Master hosts the daily stand-up meetings and tries to avoid barriers for the team during the process. Therefore, an important skill is the ability to communicate with multiple people without getting involved with the content itself. Furthermore, he should be a facilitator and when the team is dependent on someone outside the team, he should take care of that.

The Product Owner is the person who is closest to the end-user. A challenge for this role in a data science project is the management of expectations regarding the increments and then especially the demanded reliability of the end product. For example, the customer could ask for a 100% reliability of the predictive models. However, this is almost impossible in practice and the customer should be aware that a lesser percentage could be sufficient as well, depending on the application domain. Therefore, the gap between the reliability the team could offer and what customer needs should be managed. However, it is not only needed to manage and inform the customer, but the customer should also be aware that he has to find out what reliability they require.

According to the respondents, the roles required in the Development Team can vary. They need; a person who knows what data you can provide, a person with statistical knowledge a person knowledgeable about programming, a person who can arrange things from the more technical side, and someone from the business side.

Furthermore, the respondents indicated that the Development Team in data science should have smaller team size than traditional software development teams of nine people. A high amount of people working in the team causes that the secure environment is lost. Moreover, a higher amount of data scientist working on the same topic means more problems in sharing insights. However, having at least two data scientists is useful as they can check and assist each other. Furthermore, persons with different roles that support the data scientist are required.

VII. Refined Scrum-DS

In this section, the improved design for Scrum-DS is presented. By evaluating the respondents' opinions on the compatibility of Scrum-DS, we were able to discover the use of Scrum in CRISP-DM. All respondents agreed that Scrum-DS allows the team to adjust the work in progress dynamically because it enables frequent interactions among team members and provide regular feedback loops from end-users. Furthermore, the elements; user stories, product and sprint backlog, daily stand-up, sprint retrospective, sprint review refinements, and roles; were all positively evaluated by the respondents. They are a valuable and feasible addition to Scrum-DS.

However, based on the interviews some elements of Scrum were less feasible. Specifically, the use of the sprint event and increment artefact in the data preparation step. The experts indicate that combining the steps data preparation and modelling in a time-boxed sprint led to problems. The data preparation is challenging to finish in a fixed period. Consequently, it is difficult to deliver an increment with business value.

For this reason, the following change is made to Scrum-DS based on the results of the analysis. We suggest the use of a separate sprint zero for the data preparation. Sprint zero is a familiar element applied in software development [40]. It is an additional time-boxed sprint that occurs before the start of development and focuses on the collections of requirements. This helps to identify and prioritize the product backlog [41, 42]. Sprint zero in Scrum-DS will be used to prepare the data for the modelling step. During this, the team can investigate the context and identify the goals for the rest of the project. After finishing sprint zero the team has already done most of the data preparation work and can create accurate user stories for the modelling step. In comparison with the sprint during the modelling step, the sprint zero does not deliver incremental value. An overview of the refined Scrum-DS is shown in table 4. The changes with the first design are highlighted.

TABLE IV. REFINED SCRUM-DS

CRISP-DM step	Business understanding	Data understanding	Data preparation	Modelling	Evaluation
Events	• Refinement		• Sprint zero • Daily stand-up	• Sprint • Daily stand-up	• Sprint retrospective • Sprint review
Artefacts	• User stories • Product backlog • Sprint backlog			• Increment	
Roles	• Product Owner • Scrum Master • Development Team		• Scrum Master • Development Team	• Scrum Master • Development Team	• Product Owner • Scrum Master • Development Team

VIII. CONCLUSION

This study evaluates Scrum-DS, a Scrum-based data science method that combines Scrum elements with the CRISP-DM method. The design of Scrum-DS is based on three individual designs of an Agile data science method that were made by students based on a literature review. After the demonstration and evaluation of Scrum-DS in expert interviews, problems were identified. These problems overlap with typical problems when changing from a traditional process to Scrum. For example, in the beginning it is challenging but when the team gains experience with the method they get used to it . However, the problem to finishing the data preparation step in a time-boxed sprint requires extra attention in a data science project. During the data preparation step, the development team is often dependent on the availability of data. This, dependency can consume all the time left for the sprint. Consequently, it is difficult to apply the sprint event and deliver incremental value. Therefore, we improved Scrum-DS by splitting the sprint in separate sprints following the business and data understanding steps. First, sprint zero for data preparation. Second, the traditional sprint for modelling step to deliver incremental value.

From a practitioner's perspective, the results of this study are valuable as it enables practitioners in using Scrum in data science projects. The study did not apply a mixture of agile methods but used a complete Scrum method. Therefore providing a compatible Scrum data science method for guiding data science projects to successful results.

There are also some limitations to take into account when using the results of this research. First of all, the lack of demonstration on a real-life project leaves room to wonder how the project method would work in a real-life data science project. Next, as three different researchers demonstrated the design during an interview in three different organizations, there may have been some bias in the expert's responses. Last, interview results were not used in subsequent interviews to check for consensus among experts. This limits validation on problems with Scrum-DS among experts.

As for future research, we plan to improve Scrum-DS by applying it in a real data science project and to reflect on the user's experience. For further evaluation, we aim to use the framework for evaluation in design science (FEDS). The FEDS is introduced alongside a process to guide researchers in evaluating the artefacts that were designed during DSR projects [43]. This research did a first round of evaluation in an artificial context by interviewing experts on their expectations of the designed artefact. This led to a formative evaluation to improve the design for later evaluations. Future research will have a more naturalistic and summative evaluation to extend the quick and simple evaluation strategy of FEDS used in this research.

ACKNOWLEDGEMENTS

We thank Hüseyin Sener, Mariska Blasweiler, and Maryam Donker-Rostamy for their efforts in data collection and valuable contribution towards this research.

REFERENCES

[1] V. Grover, R. H. L. Chiang, T. Liang, and D. Zhang, "Creating Strategic Business Value from Big Data Analytics : A Research Framework," *J. Manag. Inf. Syst.*, vol. 35, no. 2, pp. 388–423, 2018.

[2] J. Walker, "Big data strategies disappoint with 85 percent failure rate.," *Digital Journal.*, 2017. [Online]. Available: http://www.digitaljournal.com/tech-and-science/technology/big-data-strategies-disappoint-with-85-percent-failure-rate/article/508325. [Accessed: 11-Feb-2020].

[3] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *Knowl. Eng. Rev.*, vol. 25, no. 2, pp. 137–166, 2010.

[4] J. S. Saltz, D. Wild, N. Hotz, and K. Stirling, "Exploring Project Management Methodologies Used Within Data Science Teams," in *Twenty-fourth Americas Conference on Information Systems, New Orleans, 2018*, 2018, pp. 1–5.

[5] J. S. Saltz, "The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness," in *Proceedings - 2015 IEEE International Conference on Big Data*, 2015, pp. 2066–2071.

[6] D. Larson and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science," *Int. J. Inf. Manage.*, vol. 36, no. 5, pp. 700–710, 2016.

[7] C. Dremel, M. M. Herterich, J. Wulf, J.-C. Waizmann, and W. Brenner, "How Audi AG established big data analytics in its digital transformation," *MIS Q. Exec.*, vol. 16, no. 2, pp. 81–100, 2017.

[8] J. Baijens and R. W. Helms, "Developments in Knowledge Discovery Processes and Methodologies : Anything New ?," in *Twenty-fifth Americas Conference on Information Systems*, 2019, pp. 1–10.

[9] L. Williams, "Agile Software Development Methodologies and Practices," *Adv. Comput.*, vol. 80,

pp. 1–44, 2010.

[10] C. Schmidt and W. N. Sun, “Synthesizing Agile and Knowledge Discovery: Case Study Results,” *J. Comput. Inf. Syst.*, vol. 58, no. 2, pp. 142–150, 2018.

[11] G. S. do Nascimento and A. A. de Oliveira, “An Agile Knowledge Discovery in Databases Software Process,” in *The Second International Conference on Advances in Information Mining and Management compliance*, 2012, pp. 343–351.

[12] N. W. Grady, J. A. Payne, and H. Parker, “Agile big data analytics: AnalyticsOps for data science,” in *Proceedings 2017 IEEE International Conference on Big Data*, 2017, pp. 2331–2339.

[13] J. S. Saltz, R. Heckman, and I. Shamshurin, “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.

[14] J. S. Saltz, I. Shamshurin, and K. Crowston, “Comparing Data Science Project Management Methodologies via a Controlled Experiment,” in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017, pp. 1013–1022.

[15] S. Ahangama and D. C. C. Poo, “What methodological attributes are essential for novice users to analytics? - an empirical study,” in *International Conference on Human Interface and the Management of Information*, 2015, pp. 77–88.

[16] J. S. Saltz, “Identifying the key drivers for teams to use a data science process methodology,” in *26th European Conference on Information Systems*, 2018.

[17] F. K. Y. Chan and J. Y. L. Thong, “Acceptance of agile methodologies : A critical review and conceptual framework,” *Decis. Support Syst.*, vol. 46, no. 4, pp. 803–814, 2009.

[18] C. Dremel, I. Management, and S. Gallen, “Actualizing Big Data Analytics Affordances : A Revelatory Case Study,” *Inf. Manag.*, 2018.

[19] K. Schwaber and J. Sutherland, “The Scrum Guide: The Definitive The Rules of the Game,” 2017.

[20] M. Muntean and T. Surcel, “Agile BI – The Future of BI,” *Inform. Econ.*, vol. 17, no. 3, pp. 114–124, 2013.

[21] B. M. Félix, E. Tavares, and N. W. F. Cavalcante, “Critical success factors for Big Data adoption in the virtual retail: Magazine Luiza case study,” *Rev. Bus. Manag.*, vol. 20, no. 1, pp. 112–126, 2018.

[22] L. Rao, M. MCNaughton, and G. Mansingh, “An Agile Integrated Methodology for Strategic Business Intelligence (AimS-BI),” in *Twenty-fourth Americas Conference on Information Systems*, 2018, pp. 1–10.

[23] J. S. Saltz and A. Sutherland, “SKI : An Agile Framework for Data Science,” in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 3468–3476.

[24] J. S. Saltz and A. Sutherland, “SKI : A New Agile Framework that supports DevOps , Continuous Delivery , and Lean Hypothesis Testing,” in *Proceedings of the 53rd Hawaii International Conference on System Sciences.*, 2020.

[25] W. A. Günther, M. H. R. Mehrizi, M. Huysman, and F. Feldberg, “Debating big data: A literature review on realizing value from big data,” *J. Strateg. Inf. Syst.*, vol. 26, no. 3, pp. 191–209, 2017.

[26] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Mag.*, vol. 17, no. 3, p. 37, 1996.

[27] P. Chapman *et al.*, “Crisp-Dm 1.0,” 2000.

[28] J. S. Saltz and I. Shamshurin, “Big Data Team Process Methodologies : A Literature Review and the Identification of Key Factors for a Project ’ s Success,” in *Proceedings - IEEE International Conference on Big Data*, 2016, pp. 2872–2879.

[29] Y. Li, M. A. Thomas, and K.-M. Osei-Bryson, “A snail shell process model for knowledge discovery via data analytics,” *Decis. Support Syst.*, vol. 91, pp. 1–12, 2016.

[30] S. Ahangama and D. C. C. Poo, “Designing a Process Model for Health Analytic Projects,” in *PACIS 2015 Proceedings. 3.*, 2015.

[31] A. R. Hevner, S. T. March, J. Park, S. Ram, and S. Ram, “Design Science in Information Systems Research,” *MIS Q.*, vol. 28, no. 1, pp. 75–105, 2004.

[32] R. J. Wieringa, *Design Science Methodology for information systems and software engineering*. Springer Berlin Heidelberg, 2014.

[33] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[34] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*. Pearson Education LTD, 2009.

[35] R. K. Yin, “Robert K . Yin . (2014). Case Study Research Design and Methods (5th ed .). Thousand Oaks , CA : Sage . 282 pages .,” *Can. J. Progr. Eval.*, no. March 2016, pp. 1–5, 2018.

[36] S. R. Dharmapal and K. Sikamani Thirunadana, “Big data analytics using agile model,” in *International Conference on Electrical, Electronics, and Optimization Techniques*, 2016, pp. 1088–1091.

[37] M. Das, R. Cui, D. R. Campbell, G. Agrawal, and R. Ramnath, “Towards methods for systematic research on big data,” *Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, pp. 2072–2081, 2015.

[38] M. McNaughton, L. Rao, and G. Mansingh, “An agile approach for academic analytics: a case study,” *J. Enterp. Inf. Manag.*, vol. 30, no. 5, pp. 701–722, 2017.

[39] R. Schüritz, E. Brand, G. Satzger, and J. Bischhoffshausen, “How To Cultivate Analytics Capabilities Within an Organization ? – Design and Types of Analytics Competency Centers,” in *Proceedings of the 25th European Conference on Information Systems (ECIS)*, 2017, pp. 1–15.

[40] M. R. J. Qureshi, A. Barnawi, and A. Ahmad, “Proposal of Implicit Coordination Model for Performance Enhancement Using Sprint Zero,” *I.J. Inf. Technol. Comput. Sci.*, vol. 9, pp. 45–52, 2012.

[41] M. Najafi and U. E. Engineer, “Two Case Studies of User Experience Design and Agile Development,” in *Agile 2008 Conference*, 2008, pp. 531–536.

[42] C. R. Jakobsen and K. A. Johnson, “Mature agile

with a twist of CMMI Mature Agile with a twist of CMMI," in *Agile 2008 Conference*, 2008, pp. 212–217.

[43] J. Venable, J. Pries-Heje, and R. Baskerville, "FEDS: A Framework for Evaluation in Design Science Research," *Eur. J. Inf. Syst.*, vol. 25, no. 1, pp. 77–89, 2016.