Data Analytics Project Methodologies: Which one to choose?

DOI:
10.1145/3437075.3437087

Document status and date:
Published: 15/05/2020

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Data Analytics Project Methodologies: Which one to choose?

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

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

Rob Kusters
Department of Information Science
Open University
Heerlen, The Netherlands
rob.kusters@ou.nl

Abstract

Developments in big data have led to an increase in data analytics projects conducted by organizations. Such projects aim to create value by improving decision making or enhancing business processes. However, many data analytics projects still fail to deliver the expected value. The use of process models or methodologies is recommended to increase the success rate of these projects. Nevertheless, organizations are hardly using them because they are considered too rigid and hard to implement. The existing methodologies often do not fit the specific project characteristics. Therefore, this research suggests grouping different project characteristics to identify the most appropriate project methodology for a specific type of project. More specifically, this research provides a structured description that helps to determine what type of project methodology works for different types of data analytics projects. The results of six different case studies show that continuous projects would benefit from an iterative methodology.

CCS Concepts

• Information systems→ Data analytics

Keywords

Data Analytics; Project characteristics; Project Methodologies

1. Introduction

Modern technologies allow organizations to generate collect and store big data. By applying data analytics this data provides opportunities for organizations and leads to increased firm performance [1, 2]. Data analytics is often practised in an organization through conducting projects. In these projects, data is turned to insights to support decision making or used to create a smart solution that improves business processes. To guide these projects, process models or project methodologies are recommended in the literature [3].

Within the field of process models and project methodologies, the CRISP-DM process model is the most well-known. It provides a fairly linear way to conduct a data analytics project and describes the tasks that need to be completed to finish a project [4]. A different approach, i.e. more iterative approach, is applying agile methodologies like Scrum or Kanban [5–7]. Agile methodologies originate from the software engineering discipline and provides organizations with an iterative and flexible way to conduct data analytics projects [8].

According the literature, using a process model or methodology results in higher quality outcomes and avoids numerous problems that decrease the risk of failure in data analytics projects [3]. Some problems these projects have to deal with are slow information sharing, delivering the wrong result, lack of reproducibility and inefficiencies [9, 10]. Despite that multiple methodologies are offered, a recent survey revealed that practitioners in data analytics projects merely use one, i.e. CRISP-DM. Furthermore, around 82% of data analytics practitioners do not use any data analytics methodology [11].

The existing methodologies often do not fit the characteristics of the type of data analytics project, which can be characterized in multiple ways [12]. One of them is the motivation for a project. On the one hand, a project can be driven by data and has no clear problem and the organization wants to explore what value lies in their data. On the other hand, there could be a defined problem at the start of a project and a clear solution to deliver. Another characterization of a project type is the deployment of its outcome. In some projects the outcome might have a single use, e.g. to support decision making. While the outcome of other projects is used multiple times, e.g. an algorithm to predict customer churn [13].

These different characterizations make it challenging to decide what methodology or process model to use for a specific project. Therefore, the objective of this research is to investigate what project process model or methodology is appropriate for a specific type of project. This enables organizations to improve their ability to execute data analytics projects and understand the challenges for their particular project and the process model or methodology that best mitigates those risks. For this, we formulated the following research question: *How can different data analytics project methodologies support the execution of different types of data analytics projects?*

The result of this research help organizations to increase successful investments in data analytics projects as it provides more guidance to practitioners and contributes to the professionalization of the data analytics discipline. Moreover, it helps practitioners to adopt a formal data analytics methodology. Furthermore, the research clarifies and enriches the literature on the use of data analytics process models and methodologies.

The remainder of this paper is structured as follows. Section 2 presents the theoretical background on data analytics methodologies and data analytics project types. Then, section 3

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

describes the methodology of our study. Thereafter, section 4 presents the results. Finally, a discussion and conclusion are presented in section 5 and 6, including implications to science and industry and suggestions for future research.

2. Theoretical Background

This section first reveals the five dominant methodologies to run data analytics project as shown in table 1. Thereafter, it provides an explanation on two characteristics for data analytics projects as shown in table 2.

2.1 Data analytics process models and methodologies

Finishing a data analytics project requires multiple activities that have to be completed e.g. data collection, preparation, analysing and deployment [14]. Running a data analytics project in an ad-hoc fashion results in less structure and overview on the specific status of these activities [11]. As a result, they do not retrieve the full potential of their analytics activities. Process models and methodologies provide guidelines for conducting data analytics activities. In contrast to working ad-hoc, process models and methodologies support a structured and controlled way of conducting data analytics projects. Research in process models for doing data analytics is started in the late 1990s with the Knowledge Discovery in Databases (KDD) model. This model was more focused on the data mining aspect. These initial models had a sequential nature consisting of five steps: data selection, data pre-processing, data transformation, data mining, and data interpretation/evaluation [15].

Table 1. Data Analytics Methodologies

Methodologies
Ad-hoc
Conventional
Iterative
Scrum
Kanban

After the KDD model, many other models and methodologies have been proposed [3, 5]. Similar to the original KDD model, the majority of these process models use a linear approach to completing steps and tasks defined by the methodology. Therefore, these process models are regarded as conventional methodologies. The most well-known process model is the CRISP-DM model and was developed by a consortium consisting of industry and academic representatives [16]. Although CRISP-DM was intended to be an iterative model, evidence suggests it has been used mainly in a linear fashion where a project is conducted by going through a sequence of steps [3, 4]. The model provides a set of six steps, each consisting of a number of tasks, which need to be performed to deliver value [3]. First, the Business Understanding step ensures a clear understanding of the business objectives and requirements regarding the project. Second, the Data Understanding step is to get familiar with the data, find first insights and spot data quality problems [3, 16]. Third, the Data Preparation step covers all the tasks that are related to constructing the final data set that is input for the analysis in the next step. Fourth, in the Modelling step, the right modelling technique is chosen, e.g. regression, clustering or deep learning, and applied on the prepped data set [3, 16]. Fifth, the Evaluation step ensures there is a detailed evaluation of the model to verify if the outcome meets the business objectives which were formulated in the Business Understanding step [16]. Finally, in the Deployment step, the developed model is deployed in the organization [16].

Despite the detailed description, CRISP-DM is not the solution to all managerial barriers related to data analytics. In more recent publications, new conventional models created improved versions of CRISP-DM by adding steps or tasks (e.g. problem formulation, maintenance). These provided further explanation in the activities that are needed in the specific steps [5, 6, 13, 17–19]. These new process models were introduced to cope with the specific challenges in different settings (e.g. healthcare).

Moreover, the popularity of an agile mind-set gained importance over the last years in data analytics [4, 5]. This mind-set led to the development of more flexible methods with increased focus on communication and an iterative approach. These models allow for more iteration between steps and a less sequential approach of running a data analytics project [13]. Added feedback loops provide a way to iterate the process and to create an improved output [20]. While the traditional CRISP-DM only provide feedback loops toward the business understanding after the data understanding and evaluation step, some models proposed feedback loops from different steps [21]. For example, other models distinguish two main cycles of iteration [18, 22]. One between the domain understanding, data understanding and conceptualization and the other between data preparation, modelling and evaluation. Furthermore, some models propose loops across all steps, to promote iteration [13, 17].

Furthermore, recent studies also showed the application of existing agile methods for doing data analytics projects [17, 23, 24]. The use of agile methods is common in software development. It is used because it facilitates volatile requirements and allows to quickly react to changing environments [17, 23]. Agile methods applied in data analytics consist of Scrum and Kanban [24, 25]. Scrum is an iterative process with defined events, artefact and roles to deliver value in time-boxed sprints [26]. 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 not allowed to implement suggestions for improvements on the planned work. The suggestions that arise during project execution are saved for the next sprint. Previous studies applied different elements of the Scrum method in data analytics projects. For example, in one study a method is created where all data science activities are executed in a sprint to deliver incremental value within a specific period [23, 27]. One study combined KDD and CRISP-DM as process models and added elements of Scrum [17]. For example, they used user stories to ensure that the end-user can influence the development of the end product. Furthermore, they also made use of daily stand-up meetings and sprints. Another study evaluated a design of a Scrum data analytics model. The design consisted of Scrum artefacts, events and roles that were applied on CRISP-DM [24]. Next, there is the Kanban method. The Kanban method makes use of a “Kanban board” which shows the work to do [28]. 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 simultaneously and leads to fewer problems during the process [29].

2.2 Data analytics project characteristics

Various literature identified characteristic to define data analytics project types e.g. data types, team set-up, or type of analysis [12], [30–32]. However, only two are identified that seem to influence the choice for the methodology. Firstly the way the project is driven. Secondly the deployment of the project outcome. Each of them will be discussed in the following section.

Table 2. Data Analytics Project Characteristics

Characteristics	Types
The way the project is driven	Solution driven
	Problem driven
	Data driven
Deployment of the project outcome	Single use
	Continuous use

The motivation for a data analytics project can range from well-defined to ill-defined [29, 31]. This characteristic is more relevant at the start of the project. In this paper, the way a project is driven is divided in three categories: solution driven, problem driven and data driven.

First, solution driven projects have a clear understanding of the problem that they aim to solve. The team is already familiar with the work required to finish the project. Also, the team is experienced with the data they are using [3]. Such project typically answer business questions requested by management. For this, they often use supervised methods like classification and regression [33]. The delivered models are applied in business processes and delivered as a service. The clear problem statement and focus on data modelling and deployment allow for flexible management of the project as task estimation is more accurate [30].

Second, in the problem driven project the team has a clear problem but no clear view on how to deliver the solution. The business informs the team on the problem and the team has an idea about the solution they need to create. However, they have not decided on the approach to realize the solution and they are open to different possibilities [12, 29]. In these projects, a more accurate definition of the problem and the business goals is often necessary [34]. To come to a solution they can link data analytics results to business goals, search for opportunities to turn the value of the data into a service, or discover new and valuable sources of data related to the business problem [30].

Finally, in data driven projects the data analytics practitioners have a carte blanche to find new knowledge in the data. This new knowledge can be found in the form of patterns or relations between one or more variables, represented by the data [35]. In these projects the data has a central position at the start. These are often the more advanced data science and machine learning projects. The explorative nature of such a project is considered high. The goal of the projects is to find something in the data, without knowing if this will be of value to the organization. For this, they use unsupervised methodologies as clustering and profiling and apply it on a data set [33]. Data driven projects can use data to find new business goals (goal exploration). They can search what insights might be extracted from the data (data value exploration) and by using visuals they can extract valuable stories from data (narrative exploration) [30].

The characteristic, deployment of the outcome for a data analytics project is less prominent in the literature. This characteristic represents the frequency the project outcome is deployed. This is crucial to the methodology as the result of these projects can be handled in different ways to finish a project. [5, 12, 29, 30]. In contrast to the previous described characteristic, this one is more relevant at the end of the project. In this paper, the deployment of the project outcome is divided in two categories: single use and continuous use.

First, single use projects are characterized by having a specific end and a shorter development cycle. The team is together for a limited time. A single use project is finished when the time limit is reached, or the objective is fulfilled. These projects can deliver new innovative ideas that can initiate projects that are business focused, insight report on a wide range of topics, and quick information request for a specific business question [1, 32, 36].

Second, the goal of a continuous use projects is to create, develop and support products or services that support a business process. These projects have a longer development cycles and no defined end. An ongoing flow of data needs to be analysed and the process needs to be automated and maintained [3, 13, 18]. The aim of these projects is to develop data products like dashboards or smart solutions to support business processes [1].

3. Methodology

This research aims to discover what data analytics project methodologies are appropriate for specific types of data analytics projects. As a first step, the previous section presented an overview of project methodologies and project types based on a review of the data analytics literature. The next step is to analyse the used methodologies for specific project types by collecting empirical evidence. A useful method for this is a case study since it allows for exploring and observing a new phenomenon, such as data analytics project methodologies, in a real-life context [37, 38]. Furthermore, it allows a more in-depth qualitative analysis to gain more understanding of the data analytics methodologies in their context. More specifically, we choose to apply a multiple embedded case strategy as it enables to contrast several units and to compare findings from the different case studies. To select cases we used convenience sampling because the aim of the research is a first exploration of the topic.

3.1 Data collection

A total of six case organizations were selected to be included in this research. The main criterion for selecting the case organizations was that the organization invested in data analytics to improve their business results by conducting data analytics projects. In these organizations the focus is on the different project characteristics and how they manage the project itself. Therefore, they provided multiple mini-cases that consist of different combinations of project characteristics. To obtain the required case organizations a thesis topic was formulated for master students. Data was collected by the students using interviews, a technique commonly used for data collection in case studies [41, 42]. Selection of respondents was based on their involvement in data analytics activities. More specifically, we looked for respondents that were accountable for data analytics, responsible for putting it into practice, or for executing data analytics. Furthermore, respondents needed to be active in data analytics for at least one year. Respondents that meet these criteria are considered to have enough experience to understand how the organization is conducting data analytics. Each interview followed a semi-structured approach using an interview protocol consisting of a number of questions devised by the research team (consisting of the supervisor and thesis students). The interview questions were informed by the data analytics methodologies and project types described in section 2. An example of a questions is: *To what extend do you make use of a project methodology for running data analytics projects?*

In total, the students conducted 23 interviews and the number of interviews varied based on the size of each case study organization. Therefore, at case A we conducted 4 interviews, at case B 2 interviews, at case C 2 interviews, at case D 5 interviews, at case E 4 interviews and at case F we conducted 6 interviews. Each case study was conducted by a different researcher who was connected to the specific case organization. During the interviews, the researchers were guided by an interview protocol, but extending the protocol with probing and clarifying questions if deemed necessary. Interviews were held in an online setting due to the Covid-19 pandemic. The interviews took place from March 2020 until the end of May 2020 and each of the interviews lasted half an hour to one hour. All case organizations allowed us to record the interviews on tape and the students transcribed the interviews verbatim afterwards.

3.2 Data Analysis

Analysing the interview data aimed at finding empirical evidence for the data analytics methodologies and project types. To analyse the collected data, we went through a process of selective coding. For this purpose, we used a deductive approach, which allows using a theoretical framework for the analysis of qualitative data [41, 38].

The deductive approach involved the use of a priori codes to start the coding process and these codes were derived from the methodologies and project types. These codes were used for one round of coding to mark portions of the interview data that relate to a methodology or project type. In the end, the codes were summarized into more general observations per case. The lead researcher, who was not involved in the data collection, performed the coding. He used the computer assisted qualitative data analysis (caqdas) software package Nvivo 12 for the coding of the data. Afterwards, the results were discussed with the research team to resolve any issues and inconsistencies [42–44].

4. Results

This section discusses the result for every combination of project characteristic discovered in the cases. Some cases had multiple combinations of projects type and methodologies. A complete overview of the identified project type and methodologies in the cases is highlighted in table 3. Not all combinations of characteristics were identified in the cases. The combination data driven and single use was not present.

For the combination problem driven and continuous use projects six instance where identified. This combination was present among all cases. In case A1 they develop dashboards to support business processes during these projects. These dashboards need to be maintained, thus there is continuous support. For running these projects, they make use of Scrum to have quick delivery to the business. This allows them to make progress and fast responding to the change of requirements. Furthermore, they make use of a Kanban board to create an overview and prioritize activities. Similar, case B1 uses Scrum and Kanban for the development of mobile apps. However, they state that they use Kanban when they experience impediments. This allows them to keep the project running and deliver outcomes. After the impediments are solved they turn back to Scrum. In case E1 they also make use of Scrum in their projects. Advantage of using Scrum is that after a couple of sprints, they discover and understand a number of requirements and improve future work. In contrast, the cases C1, D1 and F1 do not make use of Scrum for these projects. However they still use an iterative methodology that allows them to repeat steps until they deliver the quality they require. They have defined different activities that need to be done for the project. However the order of this is not decided. According to case D, this provides them with possibilities to adjust project goals and steps.

For the solution driven and continuous use projects, case D2 only had one example. This type of project delivers regular benchmarks for the business. Initially, the benchmark project started out with a very open mind-set. To realize this there has been intense communication with the customer to collect all requirements. After finishing this project they are able to provide new benchmarks and start new solution driven projects. These benchmarks requests consist of a specific request with a fixed dataset and results. After this, it was clear how the delivery of the end product was done. For this, they make use an iterative process as it provides more freedom to conduct the project.

Case B2 has an instance for an data driven and single use project. This project, they do during hack-day where they try to explore their data and come with new innovative ideas they can use to start new projects. For this project they have not a defined methodology and they work ad-hoc. For the data driven project type, only one instance was identified. According to case D, these projects are hard to realize as an organization tends to search what fits within

Table 3 Case Study Results

Case	Driven	Deployment	Methodology
A 1	Problem	Continuous	Scrum and Kanban
B 1	Problem	Continuous	Scrum and Kanban
B 2	Data	Single	Ad-hoc
C 1	Problem	Continuous	Iterative
C 2	Problem	Single	Iterative
C 3	Solution	Single	Ad-hoc
D 1	Problem	Continuous	Iterative
D 2	Solution	Continuous	Iterative
E 1	Problem	Continuous	Scrum
E 2	Solution	Single	Conventional
F 1	Problem	Continuous	Iterative

their strategy and this neglects them to discover new paths to success. However, an organization need to assess if their strategy is still valid and this leads to trying out new ideas. Some new ideas can initiate when they do not fit with the strategy. Then the question pops-up, if this idea need to be continued or does the strategy, needs to change. It is good to check whether an idea brings value and to take a different direction when it is clear that there is added value for the organization. However, changing the strategy will not happen quickly.

Case C2 has an example of a problem driven single use project. They run projects that are focused on the delivery of valuable data. In these, they receive a request from the business to explore value in data. From the business, they get an idea about the problem they want to tackle. However, they do not know what data to provide. This request is done one-off. Therefore, the case study uses an iterative process where they have the freedom to change the order of specific steps.

For solution driven and single use projects, there are two cases with an instance for this type. In case organization C3 these projects need a quick answer for an urgent business question. Therefore the case organization uses an ad hoc methodology. In these projects, data scientists are not involved but only business analysts. Everything is done for one occasion and is not a structural product. Often these projects can be answered in one day or at most a couple of weeks. However, when there are multiple requests on the same topic then there is the possibility to build a dashboard. The other case organization with this type of projects is case E2. They also experience that the business demands quick answers to their question. However, they prefer to use a conventional method. In these projects, activities are done that are well-know. Therefore, they are able to follow predefined steps to deliver the results.

5. Discussion

In this section, we aim to link the data analytics project characteristic with the methodology that is recommended during that case study. These links are used to develop the framework as shown in table 4.

Based on the observations in the different cases the use of iterative methodologies is prominent across the cases. The experienced freedom with this methodology is the main motivation for using it. An example of this freedom is choosing the order of project steps the team thinks is most appropriate. Also, they have more freedom to try things and iterate a step to improve the results. The use of the iterative Scrum method is also prominent in the cases. For the reason that, Scrum is more focused on time-boxed delivery of value to the customer. Therefore, they are more useful in continuous projects. These projects often have a backlog that is updated to keep the project on-track.

According to the case study results, Kanban is an addition to the Scrum method. The Kanban method can create an overview, helpful when impediments arise during the project. Interestingly the use of conventional methodologies is limited. Organization tend to dislike the linear processes to deliver data analytics results.

For deciding on the methodology for a specific type of data analytics projects, the deployment characteristic is most appropriate. The methodologies recommend for the continuous use projects are the iterative or the scrum method. The iterative nature of such methodologies allows teams to support the development of data products by implementing incremental improvements in different cycles. Especially Scrum is useful in continuous projects. The updated project backlog keep the project on-track. The suggested methodologies for single use projects showed multiple methodologies. The temporary nature of these projects led the case organizations to use ad hoc methodologies in data driven projects, iterative methodologies in problem driven projects, and apply conventional methodologies when the solution is defined.

The problem driven projects where the most occurring type of projects in the researched cases. Because most organizations emphasized the importance of business value for data analytics projects and projects without a business case should not be continued. These projects aim to solve a specific problem for the business but the road to creating a solution for this is quite vague and open to explore. Therefore, only iterative methodologies are proposed to give the team the freedom to refine their work when they get more experienced with the solution in the project.

The appropriate methodology for solution driven projects differs the most among the cases. However, the distinction between the deployment of the project results for these projects suggest that iterative is more useful for continuous and conventional together with ad-hoc for single use projects.

The case studies contained only one project that is purely driven on data. This makes it unable to make assumption on the preferred methodology for this characteristic. The type of project that was found in the case was an own initiative and the deliverables where rough versions of ideas that could be used for problem driven projects. The delivery of this rough version was done ad-hoc.

6. Conclusion

The motivation for this paper was to explore the use of methodologies to guide different types of data analytics project to successful results. The framework (table 4) we developed showed what project methodologies are most useful when considering the motivation of the project and the deployment of the outcome. The results indicate that the projects characteristic deployment of the outcome is import in choosing the right methodology.

From a practitioner's perspective, the results of this study are valuable as it enables practitioners in choosing the project methodology that fits the project they run. For example, practitioners could choose the methodology based on the duration of the project and their knowledge about the end solution.

Table 4. Data Analytics Project Methodology

Single use	• Ad-hoc (B2)	• Iterative (C2)	• Ad-hoc (C3) • Conventional (E2)
Continuous use		• Iterative (C1) • Iterative (D1) • Iterative (F1) • Scrum (E1) • Scrum and Kanban (A1) • Scrum and Kanban (B1)	• Iterative (D2)
	Data driven	**Problem driven**	**Solution driven**

There are also some limitations to take into account when using the results of this research. First of all, the limited amount of cases makes it difficult to generalize the results. Next, as four different researchers conducted the interviews in six different organizations, there may have been some bias in the responses of the interviews. Last, interview results were not used in subsequent interviews to check for consensus among interview participants. This limits validation on the specific methodology the organizations use.

As for future research, we plan to validate the framework with the help of more cases and test whether it is helpful for them to choose the right project methodology for the project they run. Furthermore, more research is needed for the data driven project type as they were underrepresented in our case sample.

7. ACKNOWLEDGMENTS

We thank Heleen Rijnkels, Marcel Slotboom, Erik van Ingen, and Avinash Parshadi for their efforts in data collection and valuable contribution towards this research. Furthermore, this research was supported by the Province of Limburg and the Center for Actionable Research Open University (CAROU)

8. 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] B. H. Wixom, B. Yen, and M. Relich, "Maximizing Value from Business Analytics," *MISQ Exec.*, vol. 12, no. 2, pp. 111–123, 2013.

[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 and I. Shamshurin, "Big data team process methodologies: A literature review and the identification of key factors for a project's success," in *Proceedings - 2016 IEEE International Conference on Big Data*, 2016, pp. 2872–2879.

[5] 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.

[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] 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.

[9] J. Gao, A. Koronios, and S. Selle, "Towards a process view on critical success factors in Big Data analytics projects," *2015 Am. Conf. Inf. Syst. AMCIS 2015*, pp. 1–14, 2015.

[10] H.-M. Chen, R. Schütz, R. Kazman, and F. Matthes, "How Lufthansa Capitalized on Big Data for Business Model Renovation," *MIS Q. Exec.*, vol. 16, no. 1, pp. 299–320, 2017.

[11] 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.

[12] J. Saltz, I. Shamshurin, and C. Connors, "Predicting data science sociotechnical execution challenges by categorizing data science projects," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 12, pp. 2720–2728, 2017.

[13] 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, Nov. 2016.

[14] J. Gao, A. Koronios, and S. Selle, "Towards a process view on critical success factors in Big Data analytics projects," in *2015 Americas Conference on Information Systems, AMCIS 2015*, 2015.

[15] 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.

[16] P. Chapman *et al.*, "Crisp-Dm 1.0," 2000.

[17] 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.

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

[19] S. Sharma, K.-M. Osei-Bryson, and G. M. Kasper, "Evaluation of an integrated Knowledge Discovery and Data Mining process model," *Expert Syst. Appl.*, vol. 39, no. 13, pp. 11335–11348, 2012.

[20] O. Marbán, J. Segovia, E. Menasalvas, and C. Fernández-Baizán, "Toward data mining engineering: A software engineering approach," *Inf. Syst.*, vol. 34, no. 1, pp. 87–107, 2009.

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

[22] S. Ahangama and D. C. C. Poo, "Unified Structured Process for Health Analytics," *Int. J. Medical, Heal. Biomed. Bioeng. Pharm. Eng.*, vol. 8, no. 11, pp. 768–776, 2014.

[23] 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.

[24] J. Baijens, R. Helms, and D. Iren, "Applying Scrum in Data Science Projects," in *IEEE 22nd Conference on Business Informatics (CBI)*, 2020, pp. 30–38.

[25] 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.

[26] L. Williams, "Agile Software Development Methodologies and Practices," *Adv. Comput.*, vol. 80, pp. 1–44, 2010.

[27] 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.

[28] 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.

[29] 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.

[30] F. Martínez-Plumed *et al.*, "CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories," in *IEEE Transactions on Knowledge and Data Engineering*, 2019, pp. 1–14.

[31] 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.

[32] S. Viaene and A. Van den Bunder, "The secrets to managing business analytics projects," *MIT Sloan Manag. Rev.*, vol. 53, no. 1, pp. 65–69, 2011.

[33] F. Provost and T. Fawcett, "Data Science for Business," *Book*, 2013.

[34] M. H. Jensen, P. A. Nielsen, and J. S. Persson, "Managing Big Data Analytics Projects: The Challenges of Realizing Value," *Proc. 27th Eur. Conf. Inf. Syst.*, no. May, pp. 0–15, 2019.

[35] W. Ayele, "Adapting CRISP-DM for Idea Mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 20–32, 2020.

[36] D. J. Power, C. Heavin, J. McDermott, and M. Daly, "Defining business analytics: an empirical approach," *J. Bus. Anal.*, vol. 1, no. 1, pp. 40–53, 2018.

[37] P. Darke, G. Shanks, and M. Broadbent, "Successfully completing case study research: combining rigour, relevance and pragmatism," *Inf. Syst. J.*, vol. 8, no. 4, pp. 273–289, 1998.

[38] 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.

[39] R. k Yin, *Case study research and applications: Design and methods*. SAGE Publications, 2017.

[40] J. Dul and T. Hak, *Case Study Methodology in Business Research*. 2008.

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

[42] A. L. Strauss, *Qualitative Analysis for Social Scientists*. Cambridge university press, 1987.

[43] T. J. Burant, C. Gray, E. Ndaw, V. McKinney-Keys, and G. Allen, "The Rhythms of a Teacher Research Group," *Multicult. Perspect.*, vol. 9, no. 1, pp. 10–18, 2007.

[44] J. Saldaña, *The coding manual for qualitative researchers*. Sage, 2015.