

A multi view approach on data analytics

Citation for published version (APA):

Baijens, J. (2021). *A multi view approach on data analytics: A process and governance perspective*. Open Universiteit.

Document status and date:

Published: 03/12/2021

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

<https://www.ou.nl/taverne-agreement>

Take down policy

If you believe that this document breaches copyright please contact us at:

pure-support@ou.nl

providing details and we will investigate your claim.

Downloaded from <https://research.ou.nl/> on date: 02 Jul. 2022

Open Universiteit
www.ou.nl



A multi view approach on data analytics

A PROCESS AND GOVERNANCE PERSPECTIVE

Jeroen Baijens



A multi view approach on data analytics: A process and governance perspective

Jeroen Bajens

ISBN: 978-90-9035372-2

Cover design & lay-out: Marjo van Pol

Printed by: 123printen.com

© Jeroen Baijens, 2021

All rights reserved. No part of this thesis may be reproduced, stored or transmitted in any form or by any means without prior permission of the author, or the copyright-owning journals for previously published chapters.

The research in this dissertation was supported by the Province of Limburg, The Netherlands, under grant number SAS-2020-03117

provincie limburg



A multi view approach on data analytics: A process and governance perspective

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Open Universiteit
op gezag van de rector magnificus
prof. dr. Th. J. Bastiaens
ten overstaan van een door het
College voor promoties ingestelde commissie
in het openbaar te verdedigen

op vrijdag 3 december 2021 te Heerlen
om 13:30 uur precies

door
Jeroen Bajens
geboren op 20 januari 1992 te Stein

Promotores

Prof. dr. ir. R.W. Helms, Open Universiteit

Prof. dr. R.J. Kusters, Open Universiteit

Leden beoordelingscommissie

Prof. dr. M.F.W.H.A. Janssen, Technische Universiteit Delft

Prof. dr. D.M. van Solingen, Technische Universiteit Delft

Prof. dr. G.M.H. Mertens, Open Universiteit


Prof. dr. J.J.M. Trienekens, Open Universiteit (emeritus)

Prof. dr. J.M. Versendaal, Open Universiteit

Dr. P. Mikalef, Norwegian University of Science and Technology

Contents

Chapter 1	
Research background, objectives, and approaches	7
Chapter 2	
What are the data analytics process methodologies?	27
Chapter 3	
Designing a Scrum data analytics process methodology	45
Chapter 4	
Data analytics project types and process methodologies	67
Chapter 5	
Data analytics governance framework	83
Chapter 6	
Developing a data analytics governance maturity model	107
Chapter 7	
Conclusion	137
<hr/>	
References	146
Summary	153
Samenvatting (Dutch)	155
Acknowledgements	157
About the author	159
List of publications	160
<hr/>	

The background features a soft-focus photograph of a dirt path winding through a forest of tall evergreen trees. Overlaid on this image is a faint, light-colored network diagram consisting of several circular nodes connected by thin lines. A large, solid teal arrow is positioned on the right side of the page, pointing horizontally towards the right edge.

Research background, objectives, and approaches

1.1 The potential of data

Awareness of the importance of data for organizations is increasing (Sivarajah et al., 2017). As a result, data has become the “new oil” in business (Parkins, 2017; Yi et al., 2014). Various websites and news outlets have illustrated this development by describing the ingenious use of data by successful companies. For example, streaming services like Spotify and Netflix rely on user data to make personalized recommendations for series or music. DHL uses traffic and weather data to optimize its transport routes, and Lufthansa uses flight data to predict the maintenance needs of its fleet (Howard, 2016; Jackman & Reddy, 2020; Jeske et al., 2013; Marr, 2017). Both small and large organizations have been inspired by these examples and realize that they too can benefit from using data (Delen & Ram, 2018). When used successfully, data can help improve decision-making in complex situations. Data can optimize business processes through smart applications or enrich products and services (Günther et al., 2017).

According to Delen and Ram (2018), there are three main reasons for the enormous growth in interest in extracting value from data. The first is the availability of technologies such as the ability to store big data and the development of mobile devices, and cloud computing. These permit more and better data analysis. The second reason is that businesses need to make better decisions. Intensifying globalization has amplified customer demand enormously, forcing organizations to make faster and better choices. Finally, there has been a cultural change: organizations are increasingly moving away from intuition and towards the use of facts in decision-making. These reasons underline the importance of data in the contemporary world. However, in practice, using data requires substantial effort. Data does not automatically turn into value; various activities must be undertaken to make it valuable for an organization. For example, insights that create business value by improving decision-making can result from transforming and analyzing data. Data analytics is one concept that embodies this creation of value. The next section will explain more about the phenomenon of data analytics

1.2. The concept of data analytics

The concept of data analytics is used frequently in the literature (Davenport, 2006; Power et al., 2018). Several definitions of data analytics have been advanced. Often, the difference between them lies in the degree to which they emphasize specific data analytics techniques or particular aspects of value creation. For example,

Ghasemaghaei et al. (2018) describe data analytics as follows: “a combination of processes and tools, including those based on predictive analytics, statistics, data mining, artificial intelligence, and natural language processing, often applied to large and possibly disperse datasets for gaining invaluable insights to improve firm decision making.” (p. 101). This definition highlights the combination of processes and tools neatly, but it only emphasizes the value of applying data analytics to improve decision-making. The definition is limited because data analytics can also be used to create innovative solutions (Günther et al., 2017). One broader definition characterizes data analytics as “the science of integrating heterogeneous data from diverse sources, drawing inferences, and making predictions to enable innovation, gain competitive business advantage, and help strategic decision-making” (Gudivada, 2017, p. 31). In this definition, the emphasis is on “science,” which causes confusion with the term “data science.” To account for these concerns, this dissertation defines data analytics as follows:

A combination of processes and tools that integrate and draw inferences from large and disperse sources of data to enable innovation, to gain business value, and to support (strategic) decision-making.

Data analytics is a rapidly evolving field. Consequently, many of those who write about the subject use synonyms side by side and interchangeably. The use of the terms “business analytics” and “advance analytics” instantiates this tendency (Boyd, 2011; Kasten, 2020). In addition, the term “data analytics” can refer to a specific type of data, such as text analytics (i.e., text data) or web analytics (i.e., clickstream data), depending on the context. Furthermore, concepts such as business intelligence, knowledge discovery, and data science are similar to that of data analytics (Davenport, 2006; Power et al., 2018). Although these concepts overlap partially, their focus is different. Therefore, unlike “business analytics” and “advance analytics,” they are not considered synonyms of “data analytics.” In this dissertation, the term “data analytics” will be used in line with the definition provided on the previous paragraph.

To apply data analytics, an organization must develop certain capabilities. According to many scientific studies, organizations should develop a so-called data analytics capability (Gupta & George, 2016; Mikalef, Pappas, et al., 2017; Wixom et al., 2013). A data analytics capability is a combination of different types of resources, such as data resources and human resources (Akter et al., 2016; Gupta & George, 2016; Mikalef, Pappas, et al., 2017). One of the leading data analytics capability models was developed by Gupta and George (2016). Their model posits that data analytics capability comprises a complex mix of tangible, human, and intangible resources,

as shown in Figure 1.1 Resources for Data Analytics Capability (Based on Gupta and George (2016)). Building data analytics capability allows organizations to create value from data and enables them to improve their business performance and to achieve a competitive edge.

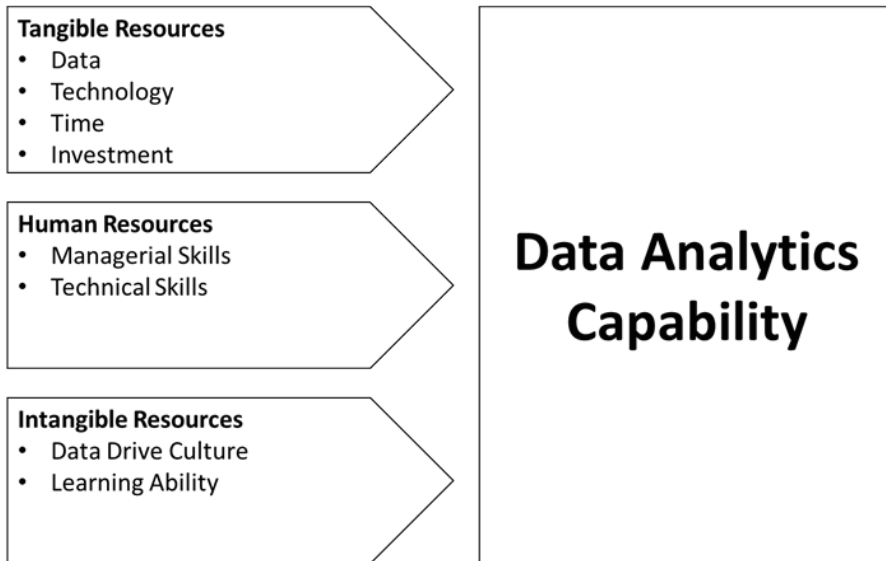


Figure 1.1 Resources for Data Analytics Capability (Based on Gupta and George (2016))

The mainstream literature typically distinguishes between different levels of data analytics, ranging from the basic to the complex. There can be as many as three, four, or even five levels (Delen & Ram, 2018; Pearlson & Saunders, 2013; Sivarajah et al., 2017). However, the most common approach is to distinguish between three levels of analytics: descriptive analytics, predictive analytics, and prescriptive analytics (Sivarajah et al., 2017).

- Descriptive analytics is the reporting aspect of analytics. It is often dubbed the basic level of analytics. It analyzes data to make it interpretable in a specific context through techniques such as visualization. Descriptive analytics tries to answer the question “what happened?” by, for example, creating dashboards (Delen & Ram, 2018; Sivarajah et al., 2017).
- Predictive analytics analyses a set of data to identify patterns that anticipate future trends and help organizations to determine what they need to do next. It focuses on answering the question “what will happen?”, and it is more advanced

than descriptive analytics. The use of techniques like predictive modeling and machine learning is common (Delen & Ram, 2018; Sivarajah et al., 2017).

- Prescriptive analytics is considered the most advanced level of analytics, and aims to answer the question “what should be done?” (Mortenson et al., 2015). Prescriptive analytics identifies the decision that has the highest probability of bringing about a successful outcome. Therefore, its use typically entails the application of techniques such as optimization and simulation (Delen & Ram, 2018; Sivarajah et al., 2017).

Another distinction focuses on the purpose of data analytics (Martínez-Plumed et al., 2019; Rose, 2016; Saltz, Shamshurin, & Connors, 2017): the business-centric approach is pitted against the data-centric approach. The business-centric approach uses the underlying business problem as the starting point of data analytics. Conversely, the data-centric approach revolves around discovering potential value in data; exploring the data is its starting point. Various analytical techniques are applied to the data set in the hope of finding interesting patterns that might be useful in generating business value (Rose, 2016). Therefore, the data-centric approach is exploratory and open ended (Martínez-Plumed et al., 2019).

1.3. Problems with the application of data analytics

Since the potential value of data analytics gained widespread recognition, an increasing number of organizations have jumped on the data bandwagon (Gupta & George, 2016). However, investing in data analytics effectively is not easy, as shown by Seddon et al. (2017). Premising their analysis on the literature, they identified sixteen different models of the realization of business value through data analytics. Each of these models revealed different factors that influence the success of an investment. The number of factors evinces the complexity of managerial decisions that target the extraction of business value from data analytics (Seddon et al., 2017). As a result, many projects do not bring about the anticipated results. So much is demonstrated by the failure rate reported by Walker (2017): 85% of data analytics projects fail to deliver the expected value. Gartner predicts that this tendency will not change considerably in the near future, and they estimate that only 20% of projects will be successful (White, 2019). Another study by McShea et al. (2016) shows that only a third of companies that invest in data analytics meet their long-term goals. All these studies demonstrate that organizations are still far from developing a successful data analytics capability that creates value for them sustainably.

These observations indicate that the value-creation potential of data analytics is not being realized. The multiple challenges that firms encounter when applying data analytics across their organizations are an important reason for this failure (Abbasi et al., 2016; George et al., 2014; Grover et al., 2018; Günther et al., 2017; Mortenson et al., 2015; Sivarajah et al., 2017). These challenges hinder the successful application of data analytics in many ways. The literature indicates that the problems tend to concern data, technology, people, and organizations (Abbasi et al., 2016; Espinosa & Armour, 2016; George et al., 2014; Grover et al., 2018; Günther et al., 2017; Saltz & Shamshurin, 2016; Sivarajah et al., 2017). In particular, the organizational problems that surround data analytics are frequently considered a major impediment (Espinosa & Armour, 2016; George et al., 2014; Günther et al., 2017). Accordingly, this dissertation will focus on organizational problems. The next section will examine these organizational problems more closely to define the scope this research.

1.4. Scope of this research

Organizational problems frequently appear in reports and studies of failures in the use of data analytics and are encountered in, among others, identifying business value, defining the scope of a project clearly, coordinating the actions of management and practitioners, establishing a data-driven decision-making culture, or adopting a siloed approach to data analytics (Been & Davenport, 2019; Davenport et al., 2020; Gao et al., 2015; Lavallo et al., 2011). Currently, these problems remain unsolved due to lack of research (Abbasi et al., 2016; George et al., 2014; Grover et al., 2018; Günther et al., 2017; Mortenson et al., 2015; Sivarajah et al., 2017).

Given the state of the art, this dissertation attempts to contribute to the mitigation of these organizational problems. Many organizational problems require a top-down approach (Grover et al., 2018; Mortenson et al., 2015; Sivarajah et al., 2017; Vidgen et al., 2017), and existing studies address the need for integration (Avery & Cheek, 2015; Espinosa & Armour, 2016; Gröger, 2018; Yamada & Peran, 2018). Accordingly, the dissertation adopts an integrated approach by focusing on the governance perspective to contribute to mitigating these organizational problems (Grover et al., 2018; Mortenson et al., 2015; Sivarajah et al., 2017; Vidgen et al., 2017). Governance-oriented research is likely to prove valuable because the existing studies on data analytics governance only identify different subjects that governance needs to address; there is no comprehensive overview (Avery & Cheek, 2015; Espinosa & Armour, 2016; Gröger, 2018; Yamada & Peran, 2018). Furthermore, processes are essential within the governance perspective. Research in data analytics processes

is already substantially founded in process methodologies. In focusing on process methodologies (Mariscal et al., 2010; Saltz, 2015), this dissertation aims not only to alleviate organizational problems but also to align itself with that research foundation (Abbasi et al., 2016; Mariscal et al., 2010; Saltz, 2015; Seddon et al., 2017; Sivarajah et al., 2017). In theory, the governance and process perspectives are seen as fundamental approaches to organizational problems (Abbasi et al., 2016; Grover et al., 2018; Günther et al., 2017). Researching these perspectives provides a broader and more diverse view of how analytics operates in an organization. As the process perspective literature is underpinned by a substantial body of theory, it is the starting point of the dissertation. The approach is then broadened through the governance perspective. The sections that follow discuss the different perspectives in greater detail.

1.4.1. Process perspective

Data analytics use process methodologies, implicitly or explicitly. According to Seddon et al. (2017), the repetitive execution of a data analytics process is one of the fundamental drivers of extracting value from data. Therefore, existing process methodologies, such as CRISP-DM, provide a consistent method of work. Process methodologies typically outline the approach to conducting data analytics (Mariscal et al., 2010). Process methodologies are also called “process models,” “project methodologies,” or “approaches” (Mariscal et al., 2010).

Within the body of knowledge, there are different process methodologies for data analytics (Saltz et al., 2018). Mariscal et al. (2010) compared many process methodologies and showed that most follow the same kinds of steps and that the most commonly used model is CRISP-DM. This model has six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment, as shown in Figure 1.2. (Chapman et al., 2000). According to Mariscal et al. (2010), these steps are recognized across different process methodologies. Several authors claim that the steps are necessary to achieve results with data analytics (Abbasi et al., 2016; Li et al., 2016; Martínez-Plumed et al., 2019). However, a recent survey revealed that process methodologies such as CRISP-DM see little use in practice. The survey also showed that there is demand for process methodology (Saltz et al., 2018). Organizations are finding it difficult to choose a process methodology to guide their data analytics (Saltz, Shamshurin, & Connors, 2017). Consequently, extensive research in the use of process methodologies that support a structured approach is needed (Abbasi et al., 2016). Accordingly, this dissertation will focus on the process perspective to contribute to the use of process methodologies in data analytics.

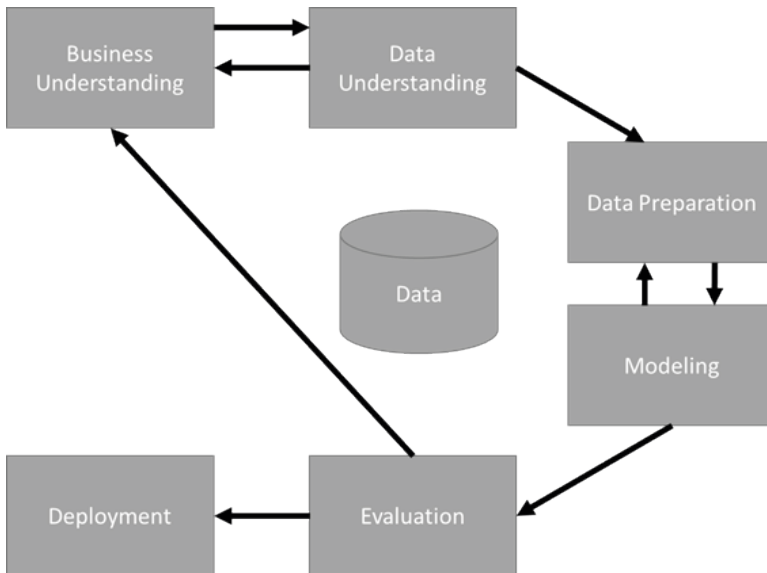


Figure 1.2. Steps of the CRISP-DM model (Based on Chapman et al. (2000))

1.4.2. Governance perspective

The term “governance” refers to the policies and practices by which a Board of Directors ensures that strategies are enacted, monitored, and executed (Rau, 2004). According to Grover et al. (2018), governance is essential for the dissemination of insights from data analysis throughout an organization, and it can catalyze value creation. Therefore, data analytics governance aims to establish policies and practices to control data analytics activities (Avery & Cheek, 2015; Gröger, 2018). The importance of data analytics governance is acknowledged by Seddon et al. (2017), who state that correct governance improves the selection of future targets for data analytics. Although the importance of data analytics governance has been stressed, there is limited research on this topic.

The literature in other fields, such as IT and data governance, shows that governance can be implemented through different mechanisms. Structural, process and relational mechanism are used most frequently (De Haes & Van Grembergen, 2004; Tallon et al., 2013). Applying the three types of mechanisms to the governance of data analytics could address numerous research gaps and problems.

The distribution of data analytics activities across organizational units is one example (Avery & Cheek, 2015; Espinosa & Armour, 2016). Another concerns the manner in

which an organization envisions the structure of its analytics function. Does it want to centralize to ensure that overall standards and protocols are applied? Or does it want to decentralize to accelerate the rate at which data analytics are deployed (Grover et al., 2018; Günther et al., 2017)? Research on the structural mechanisms of data analytics governance can provide more insights.

Furthermore, there is little coordination between managers and data analytics practitioners (Espinosa & Armour, 2016; Yamada & Peran, 2018). Therefore, data analytics must be monitored and evaluated properly to stay on track and to manage expectations. Research in the process mechanisms of data analytics governance can provide insights here. In addition, research is needed into organizational decision-making in data analytics environments and into the cultural shift from intuition-based decision-making to data-driven decision-making (Abbasi et al., 2016). Research into relational mechanisms of data analytics governance can provide more insights on this matter. Therefore, this dissertation will focus on governance perspective.

1.5. Research questions and objectives

Organizations that invest in data analytics should strive to establish a competitive advantage by developing their data analytics capability (Seddon et al., 2017). The capability in question would allow them to fulfill their business objectives and to improve their performance. However, many find it difficult to achieve these objectives owing to organizational problems. Therefore, the main purpose of this research is to contribute to the successful application of data analytics within organizations.

To attain this goal, the research covers two perspectives which were presented in the previous section: the process perspective and the governance perspective. In the process perspective, the aim is to improve the understanding of the contribution of process methodology to the success of data analytics. In the governance perspective, the aim is to improve the understanding of the role of governance in the successful application of data analytics in organizations. In order to achieve these goals, a number of research questions have been formulated. Figure 1.3. presents them, and they are discussed in detail in the next section.

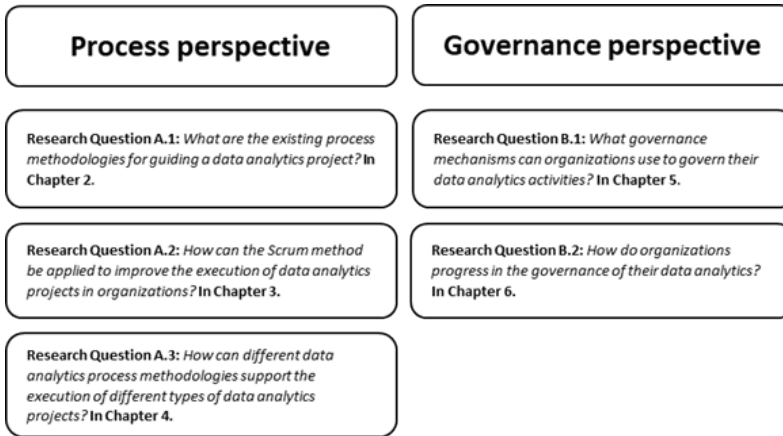


Figure 1.3. Overview Research Questions

1.5.1. Research questions: process perspective

The process perspective addresses the challenges that organizations face when they use a process methodology as a guide to their data analytics operations. The use of process methodologies has been found to result in higher-quality outcomes and to avoid numerous problems. In this way, it decreases the risk of failure in data analytics projects (Mariscal et al., 2010). Despite multiple process methodologies being presented in the literature, a recent survey revealed that 82% of practitioners do not use any of them (Saltz et al., 2018). The same survey revealed there is a demand for process methodologies. This finding suggests that there is no clear correspondence between existing process methodologies and the desires of practitioners. Organizations all have specific projects that require suitable methods (Saltz, Shamshurin, & Connors, 2017). Accordingly, this research aims to identify process methodologies that match the characteristics of different project types.

The first research objective is to investigate the state of the art in data analytics process methodology. The attainment of the objective involves a thematic overview of the relevant literature. The investigation of the process methodologies that are currently applied in data analytics projects supplies a solid basis for further research. The review will show how data analytics process methodologies have evolved and what the current state of the art is.

Research Question A.1: What are the existing process methodologies for guiding a data analytics project?

In the course of the overview of the process methodology literature, it revealed that its critics see those methodologies as too rigid. Those critics do not support the iterative and open-ended nature of data analytics projects (Saltz, 2015). Therefore, growing numbers of organizations try to apply the agile to improve the success rate of data analytics projects (Dremel et al., 2017; Larson & Chang, 2016). Previous studies argue that the agile method Scrum can achieve this end (do Nascimento & de Oliveira, 2012; Grady et al., 2017; Schmidt & Sun, 2018). Existing research tends to circle on a mixture of elements, but none apply the complete Scrum method. This leads to the second research objective, which concerns the design and validation of a Scrum-based method for data analytics projects.

Research Question A.2: How can the Scrum method be applied to improve the execution of data analytics projects in organizations?

Although agile is a helpful approach to improving the effectiveness of data analytics projects, it is not a universal solution. In particular, the results from the study in chapter 3 show that the activities that are related to the data preparation step are hard to execute in an agile manner because they are difficult to implement in an environment with time-boxed iterations. This shows that existing process methodologies often do not fit the characteristics of particular data analytics projects. Data analytics projects can be characterized in multiple ways (Saltz, Shamshurin, & Connors, 2017). The different characterizations make the selection of a process methodology challenging. Choosing an appropriate process methodology is choosing the right method for the right situation. Therefore, one of the objectives of this paper is to investigate the project characteristics that impact the choice of process methodology. This investigation will improve the understanding of the challenges and risks of particular projects. It can also improve the ability of organizations to execute data analytics projects.

Research Question A.3: How can different data analytics process methodologies support the execution of different types of data analytics projects?

1.5.2. Research questions: governance perspective

The second perspective addresses the challenges that organizations face in the of data analytics. As noted previously, data analytics provides significant opportunities to organizations. However, despite this positive influence, increasing dependence on data poses multiple issues at the organizational level, such as poor coordination between management and data analytics practitioners. While managers often aim to secure a return on investments quickly, data analytics practitioners aim at accuracy (Yamada & Peran, 2018). Another issue concerns the spread of data analytics activities across organizational units, which leads to the creation of silos. This fragmentation of efforts prevents organizations from realizing the full potential of their data analytics activities (Avery & Cheek, 2015). To address these concerns, an organization ought to implement policies and guidelines. Policies and guidelines can be created by incorporating governance. This allows that the Board of Directors ensures that activities are conducted, monitored, and achieved (Rau, 2004). For example, governance policies and practices apply to decisions and responsibilities about the management of assets (Weill & Ross, 2004).

Data analytics governance aims at the establishment of structures, policies, and controls for data analytics activities (Gröger, 2018). The notion refers to the guiding principles that are used to coordinate activities, to align interests, and to maximize the value of data analytics (Yamada & Peran, 2018). Studies of governance in other domains, such as IT and data governance, show that it is exercised through different mechanisms (Mahoney, 2018; Tallon et al., 2013; Zogaj & Bretschneider, 2014). The objective of this research is to identify the mechanisms that are relevant to creating a framework for the governance of data analytics. The framework provides concrete means to configure data analytics governance and to gauge the importance of its various aspects.

Research Question B.1: What governance mechanisms can organizations use to govern their data analytics activities?

The implementation of data analytics governance throughout an entire organization is neither quick nor straightforward. The broad scale of the different mechanisms reveals the complexity of comprehensive governance (Weber, Otto and Osterle, 2009; Tallon et al., 2013). Organizations would benefit from a sensible scheme for the implementation of concrete mechanisms, which would prevent them from committing obvious errors. In governing their analytics activities, organizations must mature gradually. Therefore, a measure for assessing data analytics governance ought to be adopted. To serve this purpose maturity models are likely to prove

helpful. A maturity model for data analytics governance would provide a useful measure of current results and areas for improvement. Thus, one objective of this research is to create, demonstrate, and evaluate a maturity model. The resultant maturity assessment instrument can be used by organizations to assess data analytics governance comprehensively.

Research Question B.2: How should organizations progress in the governance of their data analytics?

1.6. Research Methodology

This section describes the methodology of the research, which combines elements of qualitative research with design science. This approach was selected because little was known about the phenomenon under observation. The data analytics governance literature is almost non-existent, and the literature on the use of process methodologies in data analytics is limited. Therefore, rich data was needed to shed light on these phenomena. According to Edmondson and Mcmanus (2007), such situations call for a nascent theory that helps to provide tentative answers to new “how?” and “why?” questions. Often, the answers serve as a basis for further research. The opposite is true of mature theories, where much is known about existing processes and constructs. The research in this dissertation is guided by the themes and problems that emerge from the data (Edmondson & Mcmanus, 2007). The approach is flexible: it is possible to pursue promising discoveries and to abandon uninteresting results. This openness helps to ensure that key variables are identified and examined (Edmondson & Mcmanus, 2007). The dissertation answers multiple research questions. Each answer is guided by distinct methods. Table 1.1. provides an overview.

Table 1.1. Overview research methodologies

		RQ A.1	RQ A.2	RQ A.3	RQ B.1	RQ B.2
Systematic Literature Review		X				
Multiple Case Study Research				X	X	
Design Science Research	Expert Interview Evaluation		X			X
	Focus Group Evaluation					X
	Multiple Case Study Evaluation					X

1.6.1. Research Question A.1: What are the existing process methodologies for guiding a data analytics project?

Research question A.1 requires an overview of existing process methodologies for data analytics. To obtain this overview a systematic literature review was used. The review, employs an eight-step approach to avoid the omission of important contributions to the literature and to ensure that the process is detailed and transparent process (Brendel et al., 2020; Okoli & Schabram, 2012). The systematic analysis of the identified literature and the categorization of the content of the articles around data analytics process methodologies yielded a thematic overview that provides a solid basis for further research (Webster & Watson, 2002). This resulted in a scientifically rigorous systematic literature review that investigates the evolution of data analytics process methodologies.

1.6.2. Research Question A.2: How can the Scrum method be applied to improve the execution of data analytics projects in organizations?

To address research question A.2, a Scrum-based data analytics (DA) method was designed. Its purpose is to explain how Scrum would perform in a data analytics project. Therefore, design science research (DSR) was used to build and evaluate a Scrum-based DA method in practice (Hevner et al., 2004; Wieringa, 2014). Based on the DSR literature, five steps ("identify problem," "define solution," "design," "demonstrate," and evaluate) were applied. This process culminated in the development of the Scrum-based data analytics methodology (Peppers et al., 2007). The problem and the solution were defined first. Then, the literature was used to design the Scrum-based DA method. Finally, at the demonstration and evaluation stage, expert interviews were used to assess the compatibility of the method with a data analytics project. The DSR steps led to a new design and yielded knowledge about the application of Scrum to data analytics projects.

1.6.3. Research Question A.3: How can different data analytics process methodologies support the execution of different types of data analytics projects?

To answer research question A.3, it was necessary to understand what process methodology is useful for different types of data analytics projects. A multiple case study was designed to observe the process methodologies that are used in practice (Darke et al., 1998; Yin, 2017). More specifically, process methodologies were compared between organizations through a multiple embedded case study strategy.

The use of process methodologies for different type of projects within individual organizations was also analyzed (Dul & Hak, 2008; Yin, 2017). Given the importance of context, the case studies provided a view within an organization. Therefore, six organizations and 11 combinations of project types were selected to gain more insights into the selection of process methodologies and their appropriateness for different projects.

1.6.4. Research Question B.1: What governance mechanisms can organizations use to govern their data analytics activities?

A literature-based framework was created to answer research question B.1. This framework shows the mechanisms that are necessary to govern data analytics activities. It was tested in a case study to determine the extent to which it corresponds to practice (Darke et al., 1998; Yin, 2017). Context is important to improve the understanding of the governance of data analytics. The case studies enabled a contextual examination of procedures within organizations. Testing the framework in case organizations provided a deeper understanding of the novel perspective of data analytics governance. In this case, a multi-embedded case strategy was employed to identify and compare different approaches (Dul & Hak, 2008; Yin, 2017). Three different organizations were used to illustrate the implementation of data analytics governance, as such providing a deeper understanding of the possibilities to govern data analytics.

1.6.5. Research question B.2: How should organizations progress in the governance of their data analytics?

A data analytics governance maturity (DAGM) model was designed to answer research question B.2. DSR was used to build and evaluate a DAGM model (Hevner et al., 2004; Wieringa, 2014). The five steps of the DSR literature, "identify problem," "define solution," "design," "demonstrate," and "evaluate," were followed, including three cycles of design, demonstration and evaluation. The process concluded with the development of the DAGM model.

The problem and the solution were defined first. Next, three cycles of design, demonstration and evaluation were conducted. In the first cycle, the literature was used to develop the design of the DAGM model. This design was demonstrated and evaluated through interviews with experts, which yielded initial feedback on the likely practical performance of the model. The insights from the interviews were used as inputs in a redesign. In the second cycle, a focus group was used for the demonstration and evaluation stage, allowing consensus among experts to be

secured on the redesign of the DAGM model. In the third cycle, demonstration and evaluation were conducted by means of a multiple case study. The multiple case study evaluated the use of the data analytics maturity model. It supported the replication of findings across the cases and allowed the validity of the model to be gauged (Dul & Hak, 2008). It also yielded suggestions for improvement and, ultimately, a refined version.

1.7. Dissertation outline

The structure for this dissertation is as follows. Chapter 2 answers research question A1 and aims to create an understanding of data analytics 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. This research is published as a full research paper in the proceedings of the American Conference on Information System 2019.

Chapter 3 answers research question A2 by using Design Science Research it aims to understand how Scrum would integrate in a data analytics process methodology. This research is published as a full research paper in the proceedings of the Conference on Business Informatics 2020.

Chapter 4 answers research question A3 and aims to understand what type of project methodology works for different types of data analytics projects. This research is published as a full research paper in the proceedings of the International Conference on Big Data in Management 2020.

Chapter 5 answers research question B1 and aims to identify data analytics governance mechanisms to better understand how data analytics governance can be achieved. This research is published as a full research paper in the proceedings of the European Conference on Information Systems 2020 and an extended version of this chapter is published in the Journal of Business Analytics.

Chapter 6 answers research question B2 and builds an artefact to assess the maturity of data analytics governance. This research is submitted as a full research paper for the Journal of Business & Information Systems Engineering. Figure 1.4. Dissertation Outline Overview. overviews the structure of this dissertation.

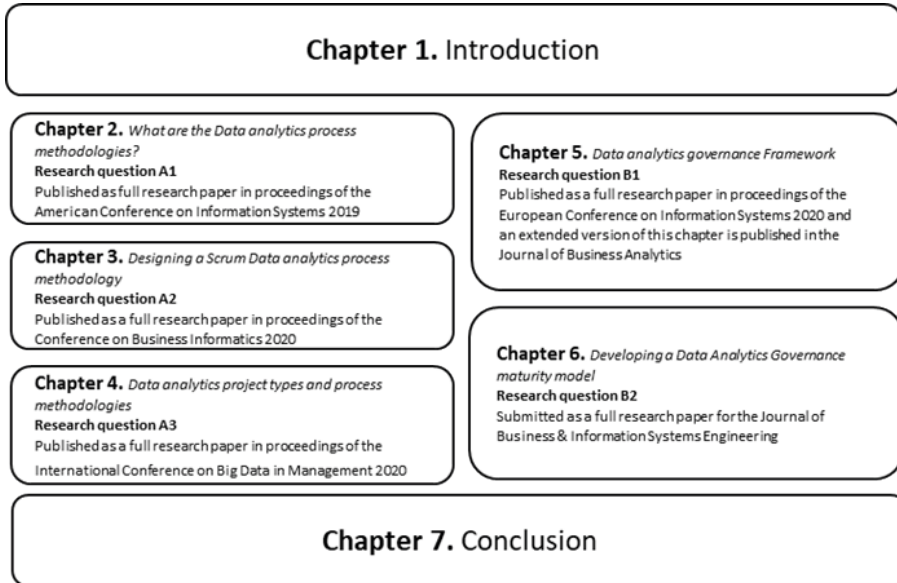


Figure 1.4. Dissertation Outline Overview

What are the data analytics
process methodologies?



The results presented in this chapter have been published as full research paper in proceedings of the American Conference on Information Systems 2019. It should be noted that in the paper the term Knowledge discovery has been used which is often used interchangeably with the term Data analytics. Furthermore, this chapter discusses different process models and methodologies. All these terms fall under the previously considered synonym 'process methodologies'.

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*

2.1. 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 & Fawcett, 2013a). 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 et al. (1996b). 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 et al. (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 et al., 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 & 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

2

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 & 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 2.2, 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.

2.2. Method

The literature review was conducted according to the guidelines presented in (Okoli & Schabram, 2012; Webster & 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.

2.2.1. 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 2.1.

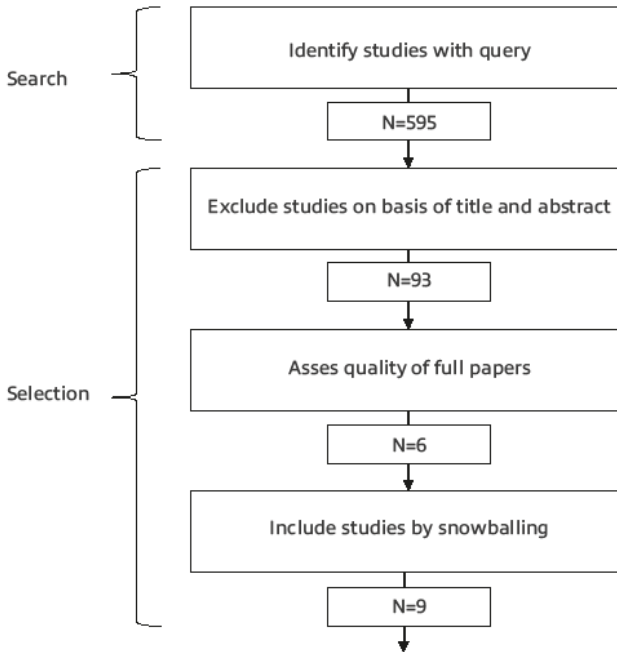


Figure 2.1. Search and selection process

2.2.2. 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 & 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).

2.3. 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.

2.3.1. Steps and tasks

Four process models in the literature (Ahangama & Poo, 2014, 2015a; Angee, 2018; Grady, 2016; Li et al., 2016) 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 2.1. 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 2.1. Overview of process model steps

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

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 (2015a) 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 (2015a) 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

2

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 *establish on-going support* step are related to the maintenance of the models. Four tasks are proposed 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 & Poo, 2015a). 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 & Poo, 2015a; Grady, 2016; Li et al., 2016). An overview of all the tasks that are added for the overall process is given in Table 2.2..

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 & Poo, 2015a). 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 & Poo, 2015a).

2.3.2. 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 & de Oliveira, 2012).

Five process models in the literature (Ahangama & Poo, 2015a; do Nascimento & de Oliveira, 2012; Grady et al., 2017; Li et al., 2016; Schmidt & 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, Shamshurin, & Crowston, 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 & Sun, 2018). In addition, *test-driven development* is an agile practice where 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 & 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 & Sun, 2018).

Table 2.2. Proposed tasks for knowledge discovery process

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

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 (do Nascimento & de Oliveira, 2012; Li et al., 2016). 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 & 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, 2016; Larson & Chang, 2016; Schmidt & 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, et al. (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, et al., 2017). An overview of the agile practices per step of CRISP-DM is given in Table 2.3.

Table 2.3. Agile practices

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

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 & Poo, 2014, 2015a) 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) and, in order to promote iteration.

2.4. 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, 2016; 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 Marbán et al. (2007). 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 & Poo, 2014, 2015a). A

well-formulated business problem will help in managing the expectation from 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 from 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 & Poo, 2015a; 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 & Poo, 2014, 2015a). 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 (2015a) 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 (2015a) 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 2.4.

Table 2.4. Existing and added tasks

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

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 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 & 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 results 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 (Ahangama & Poo, 2015a; Angee, 2018; do Nascimento & de Oliveira, 2012; Li et al., 2016; Saltz, Shamshurin, & Crowston, 2017; Schmidt & Sun, 2018).

Acknowledgements

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

The background features a soft-focus photograph of a dirt path winding through a forest of tall evergreen trees. Overlaid on this image is a faint, light-colored network diagram consisting of several circular nodes connected by thin lines. The overall color palette is muted greens and greys.

Designing a Scrum data analytics process methodology

3

The results presented in this chapter have been published as a full research paper in proceedings of the Conference on Business Informatics 2020. It should be noted that in the paper the term Data science has been used which is often used interchangeably with the term Data analytics. Furthermore, this chapter discusses different process models and methodologies. All these terms fall under the previously considered synonym 'process methodologies'.

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*

3.1. 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 (Grover et al., 2018). However, 85% of the projects that are executed fail to deliver the expected value (Walker, 2017). To guide these projects towards successful results, the use of a process model or a project methodology is recommended (Mariscal et al., 2010). 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 (Saltz et al., 2018). 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 (Saltz, 2015). Therefore, more and more organizations apply agile methods in data science projects to improve their success rate (Baijens & Helms, 2019; Dremel et al., 2017; Larson & Chang, 2016). 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 (Williams, 2010). Previous studies argue that the use of (elements of) the Scrum method improves the success rate of data science projects (do Nascimento & de Oliveira, 2012; Grady, 2016; Schmidt & Sun, 2018). 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 (Saltz, Heckman, et al., 2017). 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 (Saltz, Shamshurin, & Crowston, 2017).

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 (Chan & Thong, 2009; Saltz, 2018).

The remainder of this paper is structured as follows. Section 3.2. presents the research background on Scrum and data science process models. Next, section 3.3. presents the related work on the field of agile in data science. Then, section 3.4. describes the DSR methodology of our study. Thereafter, section 3.5. presents the design of Scrum-DS, and section 3.6. provide details on the demonstration and evaluation. In section 3.7., a refined design of the artefact is presented. Finally, a conclusion is presented in section 3.8., including implications to science and industry and suggestions for future research.

3.2. Research background

In this section, we provide background on three Scrum elements¹; artefacts, events, and roles, as shown in Table 3.1.. Furthermore, this section provides background on the most used data science process models (KDD and CRISP-DM).

¹ This study used an older version of the Scrum guide. The newest version of the Scrum guide has different terminology of the Scrum elements.

3.2.1 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 (Chan & Thong, 2009). In the end, user stories help to deliver fully realized work items in each iteration (Dremel et al., 2018). Therefore, the user story should be independent, valuable, estimable, testable, and realizable (Schwaber & Sutherland, 2017).

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 (Muntean & Surcel, 2013).

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 (Dremel et al., 2017). A user story can be put in a sprint backlog if it is small enough to be finished within one sprint (do Nascimento & de Oliveira, 2012).

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 (Félix et al., 2018). For the stakeholders, the increments are an indicator of the progress that has been made (Saltz, Shamshurin, & Crowston, 2017; Saltz & Sutherland, 2019).

3.2.2. 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 (do Nascimento & de Oliveira, 2012). 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 (Schwaber & Sutherland, 2017).

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 (Muntean & Surcel, 2013). This provides them with insights on the progress of the sprint (Saltz & Sutherland, 2019).

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 (Williams, 2010).

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 (Schmidt & Sun, 2018).

Last, the refinement happens at the beginning of a sprint where the team meets together to discuss and priorities the new user stories (Grady et al., 2017). The user stories are then combined to create a product and sprint backlog (Dremel et al., 2017).

3.2.3. 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 (Saltz & Sutherland, 2019). Fourthly, he is responsible that everyone understands Scrum (Schwaber & Sutherland, 2017). Lastly, he is also responsible that no additional items are added during a sprint (Saltz & Sutherland, 2020).

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 (Grady et al., 2017; Muntean & Surcel, 2013).

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 (Félix et al., 2018). 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.

Table 3.1. Scrum 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

3.2.4. Data science process models

To effectively engage in data science to create social or economic value, organizations have to overcome challenges at different organizational levels (Günther et al., 2017). 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 (Fayyad et al., 1996a). Further research on this model has resulted in an abundance of proposed process models and methodologies (Mariscal et al., 2010).

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 (Chapman et al., 2000). The model provides a set of six steps with tasks that need to be performed to deliver value (Mariscal et al., 2010).

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 (Chapman et al., 2000; Mariscal et al., 2010). 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 (Chapman et al., 2000; Mariscal et al., 2010). 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 (Chapman et al., 2000). 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 (Chapman et al., 2000). Despite the detailed description, CRISP-DM is not an answer to all managerial and cultural barriers related to data science.

3.3. 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 (Mariscal et al., 2010; Saltz & Shamshurin, 2016). In more recent publications, improved versions of CRISP-DM have been proposed by adding steps or tasks (e.g. problem formulation, maintenance) (Ahangama & Poo, 2015a; Baijens & Helms, 2019; Larson & Chang, 2016; Li et al., 2016; Schmidt & Sun, 2018). 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 (Li et al., 2016). 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 (Baijens & Helms, 2019). They facilitate volatile requirements and allows to quickly react to changing environments (do Nascimento & de Oliveira, 2012; Schmidt & Sun, 2018). 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 (Saltz, Heckman, et al., 2017).

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 (do Nascimento & de Oliveira, 2012; Grady et al., 2017). 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 (Schmidt & Sun, 2018). 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” (Schmidt & Sun, 2018). Furthermore, they also made use of daily stand-up meetings and sprints.

3.4. Research methodology

The research methodology chosen for this study is DSR as 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 (Hevner et al., 2004).

Design science develops artefacts that are designed to interact in a problem context (Hevner et al., 2004; Wieringa, 2014). 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 (Peppers et al., 2007):

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.

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 (Ahangama & Poo, 2015b). 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) (Saltz, 2018). This study will show how elements of Scrum are used in the CRISP-DM process model to create a method that satisfies these criteria.

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.

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.

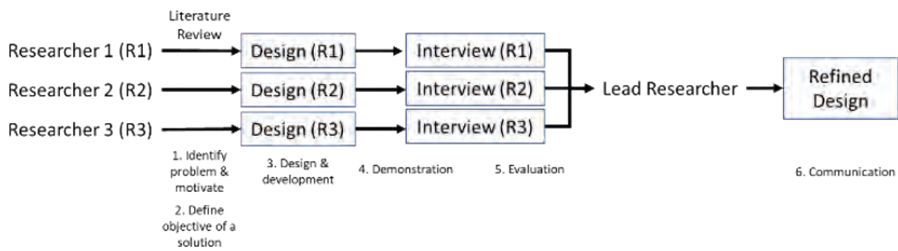


Figure 3.1. DSR steps for the development of Scrum-DS

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 (McNaughton et al., 2017; Schüritz et al., 2017).

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.

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 3.1. The execution of the different steps is provided in the following sections.

3.5. 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 3.2.

Table 3.2. Scrum-DS

	<i>Business understanding</i>	<i>Data Understanding</i>	<i>Data preparation</i>	<i>Modelling</i>	<i>Evaluation</i>
Events	<ul style="list-style-type: none"> • Refinement 		<ul style="list-style-type: none"> • Sprint • Daily stand-up 		<ul style="list-style-type: none"> • Sprint retrospective • Sprint review
Artefacts	<ul style="list-style-type: none"> • User stories • Product backlog • Sprint backlog 		<ul style="list-style-type: none"> • Increment 		
Roles	<ul style="list-style-type: none"> • Product Owner • Scrum Master • Development Team 		<ul style="list-style-type: none"> • Scrum Master • Development Team 		<ul style="list-style-type: none"> • Product Owner • Scrum Master • Development Team

3.5.1. Artefacts

The user stories in Scrum-DS can be described as an added feature to a data science product or service (Chan & Thong, 2009). 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 (Muntean & Surcel, 2013). 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 (Dremel et al., 2017). 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 (Muntean & Surcel, 2013). 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 (Dharmapal & Sikamani Thirunadana, 2016).

3.5.2. Events

In Scrum-DS a sprint of 4 weeks is preferred (Dremel et al., 2017; Schwaber & Sutherland, 2017). 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 (Dharmapal & Sikamani Thirunadana, 2016; Félix et al., 2018).

The sprint is used in combination with the activities of the data preparation and modelling steps. For the reason that work in 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 (Dremel et al., 2018). 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 (Félix et al., 2018). 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.

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.

3.5.3. 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 (Dremel et al., 2018). 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 (Schüritz et al., 2017). The Development Team is involved in every step of CRISP-DM.

3.6. 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 3.3..

3.6.1. 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.

Table 3.3. 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, 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?

3.6.2. 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.

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.

3.6.3. 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.

3.7. 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 (Qureshi et al., 2012). 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 (Jakobsen & Johnson, 2008; Najafi & Engineer, 2008). 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 3.4.. The changes with the first design are highlighted.

Table 3.4.. Refined Scrum-DS

	Business understanding	Data understanding	Data preparation	Modelling	Evaluation
Events	<ul style="list-style-type: none"> Refinement 		<ul style="list-style-type: none"> Sprint zero Daily stand-up 	<ul style="list-style-type: none"> Sprint Daily stand-up 	<ul style="list-style-type: none"> Sprint retrospective Sprint review
Artefacts	<ul style="list-style-type: none"> User stories Product backlog Sprint backlog 			<ul style="list-style-type: none"> Increment 	
Roles	<ul style="list-style-type: none"> Product Owner Scrum Master Development Team 		<ul style="list-style-type: none"> Scrum Master Development Team 	<ul style="list-style-type: none"> Scrum Master Development Team 	<ul style="list-style-type: none"> Product Owner Scrum Master Development Team

3.8. 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 (Venable et al., 2016). 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.

Data analytics project types and process methodologies

4

The results presented in this chapter have been published as a full research paper in proceedings of the International Conference on Big Data in Management 2020. This research received the award for best oral presentation. Furthermore, this chapter discusses different process models and methodologies. All these terms fall under the previously considered synonym 'process methodologies'.

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.

Keywords – *Data Analytics; Project characteristics; Project Methodologies*

4.1. 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 (Grover et al., 2018; Wixom et al., 2013). 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 (Mariscal et al., 2010).

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 (Saltz & Shamshurin, 2016). A different approach, i.e. more iterative approach, is applying agile methodologies like Scrum or Kanban (Baijens & Helms, 2019; Dremel et al., 2017; Larson & Chang, 2016). Agile methodologies originate from the software engineering discipline and provides organizations with an iterative and flexible way to conduct data analytics projects (Chan & Thong, 2009).

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 (Chen et al., 2017; Gao et al., 2015). 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 (Saltz et al., 2018).

The existing methodologies often do not fit the characteristics of the type of data analytics project, which can be characterized in multiple ways (Saltz, Shamshurin, & Connors, 2017). 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 (Li et al., 2016).

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 4.2. presents the theoretical background on data analytics methodologies and data analytics project types. Then, section 4.3. describes the methodology of our study. Thereafter, section 4.4. presents the results. Finally, a discussion and conclusion are presented in section 4.5. and 4.6., including implications to science and industry and suggestions for future research.

4.2. Theoretical background

This section first reveals the five dominant methodologies to run data analytics project as shown in Table 4.1.. Thereafter, it provides an explanation on two characteristics for data analytics projects as shown in Table 4.2..

4.2.1. Data analytics process models and methodologies

Finishing a data analytics project requires multiple activities that have to be completed e.g. data collection, preparation, analyzing and deployment (Gao et al., 2015). Running a data analytics project in an ad-hoc fashion results in less structure and overview on the specific status of these activities (Saltz et al., 2018). 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 (Fayyad et al., 1996a).

Table 4.1. Data Analytics Methodologies

Methodologies
Ad-hoc
Conventional
Iterative
Scrum
Kanban

After the KDD model, many other models and methodologies have been proposed (Baijens & Helms, 2019; Mariscal et al., 2010). 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 (Chapman et al., 2000). 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 (Mariscal et al., 2010; Saltz & Shamshurin, 2016). The model provides a set of six steps, each consisting of a number of tasks, which need to be performed to deliver value (Mariscal et al., 2010). 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 (Chapman et al., 2000; Mariscal et al., 2010). 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 (Chapman et al., 2000; Mariscal et al., 2010). 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 (Chapman et al., 2000). Finally, in the Deployment step, the developed model is deployed in the organization (Chapman et al., 2000).

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 (Ahangama & Poo, 2015a; Baijens & Helms, 2019; Larson & Chang, 2016; Li et al., 2016; Schmidt & Sun, 2018; Sharma, 2012). 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 (Baijens & Helms, 2019; Saltz & Shamshurin, 2016). 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 (Li et al., 2016). Added feedback loops provide a way to iterate the process and to create an improved output (Marbán et al., 2009). 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 (Angee, 2018). For example, other models distinguish two main cycles of iteration (Ahangama & Poo, 2014, 2015a). 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 (Li et al., 2016; Schmidt & Sun, 2018).

Furthermore, recent studies also showed the application of existing agile methods for doing data analytics projects (Baijens, Helms, & Iren, 2020; do Nascimento & de Oliveira, 2012; Schmidt & Sun, 2018). 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 (do Nascimento & de Oliveira, 2012; Schmidt & Sun, 2018). Agile methods applied in data analytics consist of Scrum and Kanban (Baijens, Helms, & Iren, 2020; Saltz, Shamshurin, & Crowston, 2017). Scrum is an iterative process with defined events, artefact and roles to deliver value in time-boxed sprints (Williams, 2010). 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 (do Nascimento & de Oliveira, 2012; Grady et al., 2017). One study combined KDD and CRISP-DM as process models and added elements of Scrum (Schmidt & Sun, 2018). 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 (Baijens, Helms, & Iren, 2020). Next, there is the Kanban method. The Kanban method makes use of a “Kanban board” which shows the work to do (Saltz & Sutherland, 2019). 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 (Saltz, Heckman, et al., 2017).

4.2.2. 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 (Das et al., 2015; Martínez-Plumed et al., 2019; Saltz, Shamshurin, & Connors, 2017; Viaene & Bunder, 2011). 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 4.2. 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 (Das et al., 2015; Saltz, Heckman, et al., 2017). 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 (Mariscal et al., 2010). Such project typically answer business questions requested by management. For

this, they often use supervised methods like classification and regression (Provost & Fawcett, 2013b). 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 (Martínez-Plumed et al., 2019).

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 (Saltz, Heckman, et al., 2017; Saltz, Shamshurin, & Connors, 2017). In these projects, a more accurate definition of the problem and the business goals is often necessary (Jensen et al., 2019). 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 (Martínez-Plumed et al., 2019).

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 (Ayele, 2020). 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 (Provost & Fawcett, 2013b). 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) (Martínez-Plumed et al., 2019).

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 (Baijens & Helms, 2019; Martínez-Plumed et al., 2019; Saltz, Heckman, et al., 2017; Saltz, Shamshurin, & Connors, 2017). 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 (Grover et al., 2018; Power et al., 2018; Viaene & Bunder, 2011).

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 analyzed and the process needs to be automated and maintained (Ahangama & Poo, 2015a; Li et al., 2016; Mariscal et al., 2010). The aim of these projects is to develop data products like dashboards or smart solutions to support business processes (Grover et al., 2018).

4.3. 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 analyze 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 (Darke et al., 1998; Yin, 2017). 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.

4.3.1. 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 (Saunders et al., 2009; Strauss, 1987). 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 extent 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.

4.3.2. 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 (Saunders et al., 2009; Yin, 2017).

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 (Burant et al., 2007; Saldaña, 2015; Strauss, 1987).

4.4. 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 4.3.. Not all combinations of characteristics were identified in the cases. The combination data driven and single use was not present.

Table 4.3. 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

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 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.

4.5. 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 4.4..

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.

Table 4.4. Data Analytics Project Methodology

Single use	<ul style="list-style-type: none"> • Ad-hoc (B2) 	<ul style="list-style-type: none"> • Iterative (C2) 	<ul style="list-style-type: none"> • Ad-hoc (C3) • Conventional (E2)
Continuous use		<ul style="list-style-type: none"> • Iterative (C1) • Iterative (D1) • Iterative (F1) • Scrum (E1) • Scrum and Kanban (A1) • Scrum and Kanban (B1) 	<ul style="list-style-type: none"> • Iterative (D2)
	Data driven	Problem driven	Solution driven

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. Organizations 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 keeps 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 were 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 suggests 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 were rough versions of ideas that could be used for problem driven projects. The delivery of this rough version was done ad-hoc.

4.6. 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 4.4.) 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.

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.

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).



Data analytics governance framework

The results presented in this chapter have been published as a full research paper in proceedings of the European Conference on Information Systems 2020. This research was nominated for the best paper award. Furthermore, an extended version of this chapter is published in the Journal of Business Analytics.

Abstract

The rise of big data has led to many new opportunities for organizations to create value from data. However, at the same time the increasing dependence on data poses many challenges for organizations in managing data analytics activities. For example, data analytics activities are fragmented across the organization resulting in incompatible outcomes. This inhibits the organization from gaining full potential of their data analytics activities.

To overcome these challenges organizations have to implement governance for their data analytics activities. IT and Data Governance literature shows that governance can be implemented through several types of governance mechanisms: structural, procedural and relational mechanisms. However, the literature is not very abundant when it comes to describing these mechanisms. Therefore, there is a need to identify data analytics governance mechanisms to better understand how data analytics governance can be achieved.

To this end, a literature review was conducted to identify a preliminary framework. The framework was validated, and extended, in three case studies by identifying practical implementations of governance mechanisms. It resulted in an extended reference framework for data analytics governance describing several structural, process and relational mechanisms. This framework can assist managers in designing data analytics governance mechanisms for their specific organization.

Keywords – *data analytics; data analytics governance; big data; knowledge discovery*

5.1. Introduction

The rise of big data technologies and advanced data analytics tools has led to new opportunities for organizations to create value from data. Analyzing their data provides organizations with new insights to improve decision making but it can also enable the creation of smart services to advance their service offerings (Davenport et al., 2012; Grover et al., 2018). Consequently, organizations are allocating an increasing number of resources to data analytics activities in an attempt to create a competitive advantage (McAfee et al., 2012; Mikalef, Pappas, et al., 2017).

Despite the fact that data analytics provides organizations with great opportunities, the increasing dependence on data poses also many challenges for organizations in managing data analytics. These challenges are more managerial and cultural in nature rather than technological, as demonstrated by two different surveys amongst executives and professionals (Lavalle et al., 2011; Wegener & Sinha, 2013). An example of a managerial challenge is the lack of alignment between management and data analytics practitioners. While management often aims for a quick return on investments, data analytics practitioners aim for accurate results (Yamada & Peran, 2018). Another example involves the spread of data analytics activities across organizational units, which leads to the creation of silos. This fragmentation of efforts prevents the organization from realizing the full potential of their data analytics activities (Avery & Cheek, 2015).

In order to address these issues, organizations have to govern their data analytic activities (Gröger, 2018). Governance is amongst others concerned with the allocation of decision rights and helps organizations in setting up procedures and policies on how data analytics activities should be conducted (Khatri & Brown, 2010). Moreover, it should protect the organization from the growing liability issues concerning data analytics activities, and should support training in the use of data analytics throughout the organization (Avery & Cheek, 2015).

Existing academic research on data analytics governance is limited and mainly addresses the need for an effective data analytics governance framework (Avery & Cheek, 2015; Espinosa & Armour, 2016; Gröger, 2018; Grover et al., 2018). There are some data analytics governance frameworks described in the practitioner's literature, but these frameworks lack empirical evidence (Oestreich, 2016). Governance frameworks in other domains, i.e. IT and data governance, show that governance is exercised through different governance mechanisms (He & Mahoney, 2006; Tallon, 2013; Zogaj & Bretschneider, 2014). Therefore, this research aims to create a governance framework for data analytics and to answer the following

research question: *What governance mechanisms can organizations use to govern their data analytics activities?*

To answer this question we conducted a literature review to develop a preliminary framework of analytics governance mechanisms and applied a multiple case study approach to evaluate and instantiate the mechanisms in this framework. In total, we conducted three case studies and collected qualitative data from 21 interviews. For analyzing the data from the interviews we applied a combined deductive and inductive coding approach. Finally resulting in an extended framework of data analytics governance mechanisms.

The remainder of this paper is structured as follows. Section 5.2. presents the theoretical background and the preliminary analytics governance framework. Then, section 5.3. describes the method of our research. Thereafter, section 5.4. presents the results from analyzing the case study data. Finally, a discussion and a conclusion is presented in section 5.5. including suggestions for future research.

5.2. Research background

5.2.1 Data analytics

Data analytics itself is defined as *“realization of business objectives through reporting of data to analyze trends, creating predictive models to foresee future problems and opportunities and analyzing/optimizing business processes to enhance organizational performance”* (Delen & Demirkan, 2013, p. 361). Techniques used for data analytics draw upon different disciplines including software engineering, statistics and machine learning (Lavalley et al., 2011). Furthermore, data analytics is considered more advanced than traditional reporting, which is mainly descriptive in nature and describes what happened. While the outcome of data analytics is more predictive and prescriptive in nature and predicts what will happen or should happen (Abbasi et al., 2016; Kiron et al., 2011). Amongst academics and practitioners, different terms are used interchangeably for data analytics and include data mining, big data analytics, business analytics, knowledge discovery and data science. In essence, all these terms refer to an activity involving analysis and exploration of data to find new and interesting patterns in data to improve decision making (Davenport, 2006). However, some terms were more often used in the past, e.g. data mining and knowledge discovery, and some accentuate a specific focus or application, e.g. big data analytics and business analytics. While data science is

considered a broader concept and refers to the scientific discipline that studies and advances data analytics.

An important development that fueled investments in data analytics is the advent of big data. This data is characterized by high volume, variety and velocity (Watson, 2014) and new technologies such as smart devices contributed to the generation of large sets of data (Chen et al., 2012). Due to big data, the size, complexity, and tools and techniques used on a dataset became critical factors (Ward & Barker, 2013). However, when handled in the right way it provides many business opportunities for organizations. Consequently, different types of organizations have been able to develop their own data analytics based improvements to remain competitive (Davenport, 2013). Furthermore, data analytics can influence organizational performance directly by improving efficiency, coordination or decision making, but also indirectly by improving the image and reputation of the organization (Grover et al., 2018).

To effectively use big data, organizations have to overcome challenges at different organizational levels in order to create social or economic value (Günther et al., 2017). To overcome these challenges, one stream of research focused on process models and methodologies, which provide guidelines for conducting data analytics projects. Research into the use of these models and methodologies started in late 1990s and has resulted in an abundance of proposed process models and methodologies (Bajens & Helms, 2019; Mariscal et al., 2010). The most well-known model is CRISP-DM and was developed by a consortium consisting of industry and academic representatives (Chapman et al., 2000). This model comprises the following steps: business understanding, data understanding, data preparation, modeling, evaluation and deployment. Despite the detailed description, process models and methodologies are not an answer to all managerial and cultural barriers related to data analytics. In the end, organizations strive to avoid all barriers and bundle all their resources to establish a big data analytics capability. In order to achieve this, governance is needed to apply policies and give strategic direction to data analytics activities (Mikalef, Krogstie, et al., 2017).

5.2.2. Governance

In general, governance refers to the rules and practices by which the board of directors ensures strategies are in place, monitored, and achieved (Rau, 2004). Governance initially starts at the corporate level where it provides a framework to support managers in their day-to-day activities (Rau, 2004). At lower levels in the organization, governance is applied to particular business domains. For example,

IT governance is focused on how firms govern physical IT artifacts (Tallon et al., 2013). While data governance focuses on governing data assets that can have potential value (Khatri & Brown, 2010). It refers to completeness of decision rights and responsibilities concerning the management of data assets. In the context of data analytics, governance is aimed at establishing structures, policies, rules and controls for data analytics activities (Gröger, 2018). Or in other words it refers to guiding principles to coordinate activities and aligning interest to maximize the value of data analytics (Yamada & Peran, 2018). In addition, a governance domain that deals with similar challenges is that of business intelligence (BI) governance (Niño et al., 2020; Watson & Wixom, 2007). However, data analytics governance is broader and includes all types of data analysis techniques (Gröger, 2018). This is illustrated by the Schüritz et al. (2017), they explain the difference in a competence center for BI that is transaction-oriented (focus on present) and a data analytics competence center that is knowledge-oriented (focus on future).

Literature in various fields shows that governance is implemented through different types of mechanisms (Alhassan et al., 2016; He & Mahoney, 2006; Otto, 2011; Tallon et al., 2013; Weber et al., 2009; Weill & Ross, 2005; Zogaj & Bretschneider, 2014). Not all studies use the same definition for each of the mechanisms, but the meaning is often very similar. For example, in IT governance, Almeida et al. (2013) describe a mechanism called *process mechanisms*, while Tallon et al. (2013) describe this as *procedural mechanisms*. In this paper, we use the following three types of mechanisms; *structural*, *process* and *relational* (Almeida et al., 2013; De Haes & Van Grembergen, 2004; Tallon, 2013; Wu et al., 2015). Structural mechanisms include mechanisms as organizational structure, roles, and responsibilities (Almeida et al., 2013; De Haes & Van Grembergen, 2004; Tallon et al., 2013; Wu et al., 2015). Process mechanisms includes mechanisms such as formal processes for ensuring daily behaviors are consistent with policies and provide input back to decisions. Different examples of these process mechanisms include routines for realization, monitoring, evaluation, and maturity of processes (Almeida et al., 2013; De Haes & Van Grembergen, 2004; Tallon et al., 2013; Wu et al., 2015). Relational mechanisms include mechanisms such as communication, participation, collaboration, education, training, shared understanding, and conflict resolution (Almeida et al., 2013; De Haes & Van Grembergen, 2004; Luo et al., 2016; Tallon et al., 2013; Wu et al., 2015).

5.2.3. Data analytics governance

A thorough review of the literature on data analytics reveals that no article explicitly mentions governance mechanisms. However, key literature on the application of data

analytics in organizations discuss issues that can lead to the creation of potential mechanisms. Therefore, this section provides a description of the mechanisms for data analytics governance according key literature. At an abstract level three different types of mechanisms can be identified which are also used in the IT and data governance literature, namely: structural, process and relational mechanisms. These mechanisms form the basis for the preliminary framework for data analytics governance as depicted in Figure 5.1. In the following sections each of the three different types of mechanisms will be elaborated to complete the framework.

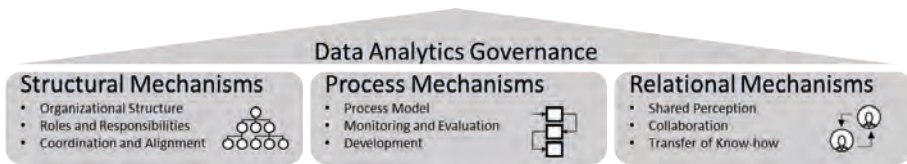


Figure 5.1. Preliminary Framework for Data Analytics Governance

Data Analytics Structural Mechanisms

The core of structural governance mechanisms for data analytics focuses on organizing data analytics functions and related decision rights. Three main themes emerged in the literature: organizational structure, roles and responsibility, and coordination and alignment (Almeida et al., 2013; De Haes & Van Grembergen, 2004; Tallon et al., 2013).

First, organizational structure embeds data analytics within the organization to understand the needs across different business units (Grossman & Siegel, 2014). The following three different organizational structures are identified. First, centralized; all data analytics activities (e.g. decision making problem prioritization) are placed in one unit. Second, decentralized; activities are spread across different units. Last hybrid; the coordination is placed in one unit and other activities are spread across different units (Grossman & Siegel, 2014). However, none of these structures is perfect and all come with certain tradeoffs. The decision for a specific structure should be based on the organization's specific context. For example, a centralized structure helps to combine activities and avoid unnecessary repetition of activities. However, this structure can provide a dependence on one unit as they own all resources, resulting in an unit that has no clear view on pressing data analytics questions at other levels (Schüritz et al., 2017).

Second, data analytics requires new diverse groups of roles with a diverse set of skills to be successful. Therefore, new roles and responsibilities need to be clearly defined to achieve a successful implementation of data analytics (Dremel et al., 2017; Grover et al., 2018). The following key roles have been observed by Schüritz et al. (2017): data scientist, project manager, data architect and business users. The different roles share responsibilities for creating leverage in the organization regarding the availability of resources, monitoring various data analytics activities, development of a data analytics platform and design of the data and information architecture for the data analytics solutions (Schüritz et al., 2017). Additionally, Kiron et al. (2011) propose to have specific roles to ensure data quality for data analytics activities.

Finally, a mechanism is required for coordination and alignment among people and organizational departments. Coordination is much needed, since data analytics activities are carried out across the organization (Espinosa & Armour, 2016). Therefore, a dedicated committee structure is proposed to promote the business value of data analytics activities to ensure that data analytics projects get the required support, but also to take care of the prioritization for projects (Dremel et al., 2018; Grossman, 2018; Grossman & Siegel, 2014; Kiron et al., 2011). This steering committee consist of executives from different departments to oversee the work of data analytics (Dremel et al., 2017). In addition, alignment of different organizational norms, values and outcomes is needed to generate business value (Kiron et al., 2014). Organizations struggle to align data analytics activities with the traditional way of decision making (Akter et al., 2016). The steering committee for data analytics can help to create alignment between data analytics and strategy by building understanding between data analytics objectives and business priorities (Akter et al., 2016).

Data Analytics Process Mechanisms

According to key literature concerning process mechanisms, governance should be used to set up routines for the realization, monitoring, evaluation, and development of analytics processes (Almeida et al., 2013; De Haes & Van Grembergen, 2004; Tallon et al., 2013). Three themes emerge for process mechanisms: process model, monitoring and evaluation, and development.

First, process models support a structured and controlled way of conducting data analytics projects. A diverse set of process models are available that lead to the deployment of data analytics results. The most well-known process model for data analytics is the CRISP-DM model, which provides a set of steps and tasks that need to be performed in order to deliver data analytics insights (Mariscal et al., 2010). A popular approach is to apply agile practices in data analytics processes (Dremel

et al., 2017). This facilitates volatile requirements and allows to quickly react to changing environments (do Nascimento & de Oliveira, 2012; Schmidt & Sun, 2018). Although a lot of organizations are not using a process model, consensus amongst academics remains that maintaining a well-defined repeatable process for data analytics projects will improve efficiency (Saltz et al., 2018).

Second, understanding and improving the level of consistency of processes requires monitoring and evaluating the efficiency and effectiveness of data analytics. Monitoring of data analytics projects enables the organization to intervene when problems arise. This ensures that data analytics efforts are leading to the desired business results (Grossman & Siegel, 2014). Furthermore, organizations need to get better in measuring the ROI from their data analytics projects and make the connection between data analytics and business outcomes (Grover et al., 2018).

Finally, a development roadmap should exist to ensure data analytics will develop towards the goals they pursue as an organization. As data analytics can consist of a diverse set of goals, organizations should have in mind how to reach these goals and how to improve on them. A maturity model is one way that can support the creation of a road map for organizations in developing their data analytics capability (Grossman, 2018).

Data Analytics Relational Mechanisms

Concerning the key literature on relational governance mechanisms, organizations should organize work in terms of interrelationships between people and groups. From this literature three main themes emerge: shared perceptions, collaboration, and transfer of knowledge and expertise (Almeida et al., 2013; De Haes & Van Grembergen, 2004; Tallon et al., 2013). First, shared perceptions on data analytics activities are crucial. For example, people in the organization should share the idea that the outcome of data analytics is often uncertain. Furthermore, organizations should keep supporting data analytics activities after the first disappointing results. This requires a strong organizational attitude that is open towards failure. While penalizing employees when they do not live up to the expected data analytics results discourages further work (Dremel et al., 2017). Organizations should provide sufficient autonomy where individuals can have their own judgment on their data analytics work (Barbour et al., 2018). Besides that, an organizational culture embracing data analytics is crucial for its success (Grover et al., 2018). It should prefer data over gut-feeling, gives room for experimentation and testing, and is open about failure, with the purpose to learn from it (Abbasi et al., 2016; Berndtsson et al., 2018; Kiron et al., 2011). Moreover, a different mind-set from management will contribute

towards more consensus and supports the team to perform less ad hoc (Yamada & Peran, 2018).

Second, the work in data analytics is perceived as multidisciplinary knowledge work and highly depends on collaboration between individuals with complementary skills (Grover et al., 2018). While previous research showed that working across disciplines provides many opportunities, it also comes with multiple challenges (Barbour et al., 2018). For instance, it increases the complexity and difficulty of managing individuals and it makes communication about data analytics work more difficult (Barbour et al., 2018). Therefore, organizations should promote communication among individuals and groups that are involved in data analytics activities. According to Dremel et al. (2018) organizations establish a social community to support employee-level collaboration by using enterprise social software to promote communication or host online meetings. Furthermore, for creating novel collaboration between data analytics stakeholders, organizations should place them close to each other in a centralized organizational unit.

Third, transfer of knowledge and expertise is crucial since organizations should ensure that they acquire and retain the right skills. The skills include: technology, modelling and analytic skills, and knowledge of the data and the business (Davenport et al., 2001). For this purpose, there should be conditions and opportunities to share know-how to learn from others (Kiron et al., 2011). Consequently, organizations should take care of the development of their employee's competencies. According to Dremel et al. (2018) they need to implement a central education program to improve data analytics skills. Now only a limited amount of organizations train employees in data analytics related disciplines (EYGM-Limited, 2015). As a result, companies are not able to get the full potential from their data. Therefore, they are forced to hire external consultants to support with data analytics work. Another option to acquire the required skills is by intense collaboration with external data analytics consultants (e.g. co-coding) to transfer knowledge (Dremel et al., 2018).

5.3. Research methodology

In this research, we aim to develop a Data Analytics Governance framework. As a first step, the previous section presented a preliminary version of the framework based on a review of the governance and data analytics literature. Next step is to evaluate the framework by collecting empirical evidence for each of the mechanisms in the framework. A useful method for evaluating our framework is a case study since it allows for exploring and observing a new phenomenon, such as data

analytics governance, in a real-life context (Darke et al., 1998). Furthermore, it allows a more in-depth qualitative analysis to foster a more holistic understanding of the data analytics governance mechanisms in their context. More specifically, we choose to apply a multiple case strategy as it supports the replication of findings across cases. Additionally, it also enables to contrast cases and to compare findings from the different case studies

5.3.1. Data collection

In total three case studies (A, B and C) were selected to be included in this research. The main criterion for selecting case study organizations was that the organization invested in data analytics to improve their business results by conducting data analytics projects. To obtain the required case organizations a thesis topic was formulated and master level students in the data science management could take part in this research during their thesis period. However, students could only apply for the topic when they had an available organization that conduct data analytics projects. At the organizations data was collected using interviews, a technique commonly used for data collection in case studies (Dul & Hak, 2008; Yin, 2017). 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. The interview questions were informed by the data analytics governance mechanisms that resulted from the literature review. Examples of the questions are *“How does the organization enable or promote the collaboration and teamwork between the people and groups who are involved in the data analytics work?”*. Before actual use of the interview protocol, the questions were tested in a pilot interview to see if anything needed clarification and if all mechanisms could be covered in roughly 60 minutes of interview.

In total, we conducted 21 interviews and the number of interviews varied based on the size of each case study organization. Therefore, at case A we conducted 13 interviews, at case B we conducted 5 interviews and at case C we conducted 3 interviews. Each case study was conducted by a different researcher, requiring upfront training of the researchers and discussing the interview protocol with them. During the interviews, each researcher was encouraged to follow the interview

protocol carefully, but extending the protocol with probing and clarifying questions if deemed necessary. Interviews were held in a personal face-to-face setting to establish trust and providing a comfortable setting for sharing data and experiences. The interviews took place from October 2018 until the end of November 2018 and each of the interviews lasted approximately one hour. Two case organizations allowed us to record the interviews (A and B) and each interview was transcribed verbatim afterwards. The third case organization (C) did not allow us to record the interview and therefore the researcher made a summary of the interviews and each of them was crosschecked with the respondent for validity purposes.

5.3.2. Data analysis

Analyzing the case study data aimed at finding empirical evidence for the mechanisms comprising the Data Analytics Governance framework. More precisely, we looked for instantiations of the governance mechanism in each of the case organizations. For this purpose we used template analysis, which allows to combine a deductive and inductive coding approach to the analysis of qualitative data (Saunders et al., 2009). This enabled us to identify concepts or main ideas hidden in the data and most likely relate to a phenomenon of interest (Saldaña, 2015).

The deductive approach involved the use of a priori codes to start the coding process and these codes were derived from the nine mechanisms in our framework. These nine codes were used in a first round of coding to mark large portions of the interview data that relate to a specific mechanism. The next round of analysis focused on the marked texts from the previous round and each governance mechanism was analyzed separately. In this round an inductive approach was used, i.e. open coding, to identify instantiations of each governance mechanism in the marked texts. This resulted in new codes that best described the instantiation that was identified. After processing all the marked text of one mechanism, the newly found codes were grouped to identify overlapping or similar codes.

One researcher, who was not involved in the data collection, performed 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 a fellow researcher to resolve any issues and inconsistencies (Burant et al., 2007; Saldaña, 2015; Strauss, 1987).

5.4. Case study results

The results of the case studies revealed that all 9 sub mechanisms were identified in at least one case organization. Moreover, there were no other mechanisms found in the cases despite asking this question to the participants consistently. A complete overview on the identified mechanisms is highlighted in Table 5.1.

5.4.1 Introducing the case organizations

Case organization A is a global biopharmaceutical company and the most advanced organization in terms of using data analytics. Their operating conditions have become increasingly challenging under the global pressures of competition. As a result, the company continually takes measures to evaluate, adapt, and improve its business practices to better meet customer needs. Therefore, they leverage digital and data capabilities across the organization. About four years ago, the company began using data analytics more purposefully. For instance, it created a central data science competency center, invested in tools and data platforms, and organized internal data science conferences. They conducted data analytics projects for creating supply chain metrics and dashboard, commercial forecasting across various markets, and optimization of equipment train setups in chemical factories. In these projects, the organization used different types of analysis including descriptive, predictive and prescriptive analysis.

Case organization B is a producer of mid-range and higher segments bicycles, and bicycle parts and accessories. They are implementing a low cost strategy to stay more competitive and to rationalize their footprint. Therefore, they focus their data analytics efforts on applications in supply chain planning (e.g. budget and demand forecast) and analyzing consumer information (e.g. using data from IoT and Social Media). The type of analytics applied is mainly descriptive and predictive in nature and status reports are used to present the outcomes to end users.

Case organization C is a professional trade association for pharmacies and their main objective is to support the promotion of medicine supply. Therefore, they collect detailed data on medicine use in the Netherlands. With data analytics, they provide regular reports (mainly descriptive) on medicine usage for their clients. Furthermore, they are experimenting in projects with more advanced predictive data analytics methods. Examples of these projects are clustering groups of prescribers, regression to predict seasonal influence, and network analytics to predict shifts between products. Despite the fact that not all of their activities are successful, the results are considered useful enough to continue the experiments.

Table 5.1. Case Comparison Data Analytics Governance Mechanisms

Categories	Data analytics governance mechanisms	Case A	Case B	Case C
Structural	Organization structure	• Hybrid	• Decentralized	• Centralized
	Coordination & alignment	• Management allocates resources • Demand management process	• Project group	• Responsible person
		• Informal discussion • Monthly formal meetings • Stand up meetings	• Informal discussion	• Informal discussion
	Roles & responsibilities	• Analytics roles: o Data engineer o Data analyst	• Analytics roles: o BI developers	• Analytics roles: o Analyst o Developers
		• Business roles: o Global process stewards o Business users	• Business roles: o Analyst (for Digital, Market, Sales, Stakeholder and, Supply chain analyst)	
		• Data science roles: o Data scientists		
• Platform roles: o IT role o Data Architect		• Platform roles: o Data Architect	• Platform roles: o IT role o Administrator	
Process	Process model	• Recognized standard innovation process • Recognized standard prototype phases	• Ad hoc	• Ad hoc
	Monitoring & evaluation	• Functionality check • Verification Business • Process track tool • Iterate quickly • Regular strategy meetings	• Human check • Protecting person	• Human check
		• Interaction end user	• Interaction end user	• Benchmarking
Development	• 2/3 year strategic roadmap • Frequent goals meetings	• Group BI persons		
Relational	Shared Perceptions	• Management support • Show how tools are used • Share success stories	• Share success stories	
	Collaboration	• Frequent team meetings • Online communication tools		• Place collaborating departments close to each other
	Transfer of Knowledge	• Personal Connections • Internal/ external conferences • Online platform • Job rotating	• Personal Connections • Short presentations	• Online platform

5.4.2. Organizational structure

The case studies revealed that the case organizations use three different structures for positioning the data analytics function in the organization. Case B has a decentralized structure and distributes its data analytics activities across the whole organization. This structure led to multiple data analytics islands and prevents them to be competitive, although they recognize that also other factors might be involved here. Alternatively, case C has a centralized structure where one department handles all data analytics activities. The centralized structure causes high dependency on this department. Finally, case A has a hybrid structure with a center of excellence on global level and multiple data analytics activities distributed across different units on functional and divisional level. They recognized three different advantages. First, the business side can request changes to analytics models by themselves instead of depending on IT, which provides them control of all different requests. Second, it provides the opportunity for sharing information and experience among different departments and help each with the necessary support or generate ideas and solutions. Last, this structure enables the business side to create their own KPI's. However, using a hybrid structure also has several drawbacks. One example is the bureaucracy that comes with it and causes delay in changes or delivery of analytical models. Another is the number of persons who get involved, making it difficult to know who is responsible for what.

5.4.3. Coordination and alignment

According to the three cases, coordination ensures that decisions on allocation of resources are based on prioritized data analytics activities. In case C the manager of the centralized unit is responsible for prioritization of analytics activities and the required resources. Similarly, in case A management decides where to allocate resources. However, the units also have some autonomy in allocating resources. For example, in the IT group they have a standard demand management process for all their IT and data analytics projects. This helps with prioritization and funding of projects. Nevertheless, for the future they aim to create a council that meets regularly to review results, take action, and identify new opportunities. In contrast, case B tries to coordinate activities from one big project group consisting of persons from multiple disciplines. This groups together decides where new smaller project teams start.

Concerning alignment, all three cases experienced that this takes place during informal communication. Although this is valuable, case A highlights that it is also problematic when persons informally discuss something without informing other

colleagues who might benefit from it. For supporting alignment, case A had two initiatives in place. First, they have monthly formal meetings with the responsible persons of different groups where they discuss what demands are going to get approved. Second, across the organization there are stand up meetings that span multiple different IT and data analytics topics. The stand-up meetings initiate different interactions among persons and groups and supports making decisions together.

5.4.4. Roles and responsibilities

The three cases revealed four main categories of roles, including their responsibilities, involved in data analytics. Within these categories the case organizations had their own set of different formalized roles. The four categories are: analytics, data science, platform and business roles. First, the analytics roles are responsible for data engineering and visualization. They use data from IT systems and put it into meaningful results based on business requirements. The analytics roles are well connected to the other role categories. For example, they push demand towards the platform role, they seek advice from the data science role in complex situations, and they work on demand of the business role. Formalized roles for analytics at case A are the data engineer and the data analyst. The data engineer transforms data to fit into a data analytics model. The data analyst focuses on turning data into something useful and presents it in an understandable way. A similar formalized role in case C is done by the analyst. Furthermore, Case C has developers that process data (ETL), conduct analyses, create programs for automatic analyses, and build and maintain results of analyses. In case B this is done by the BI developers who also perform ETL.

Second, the responsibility of the business role is to identify opportunities within the business and collaborate with stakeholders to get results. They focus on metrics and KPIs to monitor and create a demand for this. Formalized business roles in case A are business users who define requirements and definitions with regards to the requested insights. Furthermore, global process stewards ensure that these requirements are defined enterprise-wide, to avoid that activities negatively affect decision-making. In case B there are different areas for analysts: Digital , Market , Sales , Stakeholder and Supply chain analyst. Moreover, in case A and B the business roles conduct data analytics activities on their own. This is the self-service, where they can gather the required information themselves in a specific environment to make graphs from prepared data.

Third, the data science roles focus on more advanced analysis. They deliver this on request to the end user and provide support when users themselves struggle with analytics activities. In case A there is a formalized data scientist role, they transform data and harmonize it with the business, but also teach the business to do that on their own. The data scientist is more experienced using advanced data processing methods than the data analysts. In case B they hired a data scientist in the past. Unfortunately this was not a success for them, because the organization did not provide enough guidance on topics the data scientist could work on.

Last, the responsibility of the platform role is to operate and support operational systems that generate and maintain the data. They make sure that data models reflect the requirements and definitions of the business function and ensure that data is available for use. For the platform role case A formalized IT roles focus on activities from a technology perspective. These activities include: where to put the servers or which specific databases to use. Furthermore, they have a formalized data architect similar as in case B. The data architect determines how an information request can be answered, what table structure is needed, and what the details are from what they need to know. They understand the existing landscape, including projects and data artefacts that already exist. In case C formalized IT roles are responsible for delivery of data and tools that enable data analytics, but also for technical maintenance and design of their systems. Furthermore, they have an administrator who is responsible for the quality of the data and ETL procedures.

5.4.5. Process model

A formalized process model is only identified in case A. They apply an organizational standard as innovation process that they use across the organization. It provides them with a level of structure, sets expectation with stakeholders, and helps utilizing resources based on demand. Despite the fact that it is not documented it consists of the following three steps: prototype, operationalize, and industrialize.

First the prototype step uses multiple iterations to create prototypes for data analytics solutions. Depending on the complexity it can take six to twenty weeks. The phases in the prototype step are comparable to CRISP-DM. However, they state that using a method as CRISP-DM increases frustration from the business, because a step like data understanding is often a bottle neck and limits them to deliver fast results. Therefore, case A keeps the process flexible by providing the end user the possibility to change requirements during development. The prototype step consists of 5 phases: understand problem, understand data, clean and curate data, analyze, and communicate.

Once the prototype is built, the operationalizing step gives ownership to IT. The code is then verified to ensure that everything is written correctly, the formulas are valid, and there are no errors. Last, in the industrializing step, the solution is fully documented and tested. When this is done it is transferred to an operational environment to make it available to business users. From that moment there should be: a regular refresh of the data, monitoring and maintenance of the solution, and first-line support for business users.

Case B and C both did not have a standard process for doing data analytics projects. At case B they recognize a flexible method that starts with a request of a rapport which happens ad hoc. Despite the fact that case B does not have a standard process they expect it would support them to manage projects across different units. In case C they do not see value in a standard process, because the efforts to describe and maintain the process do not outweigh the benefits. However, general steps as problem understanding, understand data, clean and curate data, analyze, and communicate are recognized.

5.4.6. Monitoring and evaluation

A mechanism regarding monitoring is present in case A at the end of the prototype step. At this point the business side verifies and IT checks the functionality of the prototype. Furthermore, they have support for monitoring by a process tracking tool Jira, that is structured by their five core process phases. Moreover, they state that quick iterations during the prototype step helps them with monitoring, since it provides a visibility on the latest requirements.

According to case A monitoring is done more informal by persons themselves and happens by the power of collaboration and discussion. This discussion can happen during the regular strategy meetings where they review results on routine basis. Case B and C also rely on this personal check when data analytics results are shared among colleagues. Additionally, case B has a person responsible for protecting data analytics activities. This person is regularly the project owner or someone working in the project.

For evaluation two mechanisms were identified across the cases. First, case C include benchmarking to compare projects and evaluate how they perform. Second, case B and A favor strong interaction with the end users of insights to evaluate if the outcomes are worthy.

5.4.7. Development

To pursue development of data analytics activities case A created a strategic roadmap for the next two to three years. According to them the road of becoming a data driven organization is a long journey that includes all aspects of strategy. The multiyear roadmap is important, but also meeting the small milestones along the way. Case A is already three years into their journey and problems they face in year three are different than in year one. For example, in the beginning access to data and technology was problematic but, now this is solved and they face problems in terms of getting value from projects and finish them quickly.

In order to develop this roadmap they have frequent management meetings to discuss goals and direction to go. These goals are incorporated in day to day activities and tracked by a balanced score card to ensure they operate appropriately. In case B they grouped BI persons together to create a BI community as this was demanded in their strategy. In addition, individual efforts measured the organization in term of maturity and discovered that they were in the beginning phase and have multiple opportunities to improve upon. Still, they lack a structural plan and hence developments are uncoordinated and not connected. Consequently the goal for the future is to create a BI board that builds a strategy. In case C there is no development plan, but there are some developments implemented without roadmap.

5.4.8. Shared perceptions

Mechanisms for shared perceptions are important to establish trust in data analytics activities and create a data driven mind-set. Therefore, in case A, the right mind-set is driven from the top level of the organization. Management is aware and supportive on using data analytics by sponsorship and motivating their employees to provide a desire to move forward and give a message why it is beneficial. Furthermore, case A heavily invests in building employees trust in data to support the use of data analytics. Formerly, they experienced constant struggle with employees challenging the data and the KPIs. To change this they now show the tool to create KPIs and the data used by the tool instead of showing the result. This enables employees to start seeing the potential of the tool, and what functionality it offers for them.

To further support the trust in data analytics case A and B spread early adopters of data analytics through the organization. In order to get more persons convinced by showing success, share stories and mentoring. In addition in case C this trust grew by the years of past collaboration and the informal way of working.

5.4.9. Collaboration

According to case A collaboration is crucial as not one person has all the knowledge to solve problems in data analytics activities. Problems need to be solved from different directions with persons that have different but, complementary skills. Naturally in case A and B persons lean on each other to solve a problem or seize an opportunity.

According to case A and B communication contributes towards better collaboration and hence case A hosts frequent team meetings where they speak as a group to facilitate communication among team members. Moreover, when physical meetings are not possible they host Webex meetings to communicate in an online environment. In addition, case C purposely puts different departments close to each other. For example, the IT and the data analytics department are located next to each other to stimulate more communication.

5.4.10. Transfer of know-how

According to case A and B, transfer of know-how among persons is crucial to initiate learning and often occurs through an employee's personal connections. Therefore, case A connects persons in communities of practices by having employees regularly attend external conferences to share stories with other companies e.g. data mining seminars. In addition, the organization organizes regular data science conferences to give its employees the opportunity to share data analytics related experiences. In case B they have a similar aspiration, since they plan to organize a hackathon where employees can pitch ideas on data analytics in front of a diverse audience. Hitherto, they only organized short presentation sessions to share knowledge in an informal setting.

Furthermore, knowledge is also shared using online tools. They provide a platform to support sharing of ideas and insights. In case C they have an internal online platform to share experience on a diverse set of topics. Similarly, case A has a variety of Yammer groups that share new insights, techniques and opportunities that exist within or outside the organization.

Case A uses yet another mechanism for knowledge sharing: job rotation. This enables that data analytics practitioners to work at different places in the organization and to learn about different data analytics activities. Nevertheless, case A states that despite people are willing to learn, they do not always accept the lessons learned from other people and thus keep making the same mistakes.

5.5. Discussion and conclusion

Given the lack of research on governance mechanisms for data analytics, the primary objective of this study is to achieve a better understanding of governance mechanisms implemented by organizations to govern their data analytics efforts. As a first step, a preliminary framework is developed based on a literature review. The framework consists of two levels (see Figure 5.1.) and the first level comprises three governance mechanism categories, i.e. structural, process, and relational. At the second level, more detailed sub-mechanisms are specified in each of the three categories, resulting in a total of nine data analytics governance mechanisms. In the next step, analysis of the data from the three case studies confirmed the existence of all nine governance mechanisms (see Table 5.1.). This is demonstrated by the fact that we found at least one instantiation of each governance mechanism in one or more case studies. These instantiations are a valuable addition to our framework and are examples of how governance mechanisms can be implemented by organizations. In our analysis of the data, we also focused on mentions of mechanisms that do not fit with one of the nine governance mechanisms, but they were not found in the data collected at the three case study organizations. This suggests that the current set of mechanisms is comprehensive enough for describing data analytics governance mechanisms.

Our framework is a response to the call for action by Espinosa and Armour (2016) and Gröger (2018) who both suggest that a data analytics governance framework is needed. There have also been other responses to this call for action such as by Avery and Cheek (2015) and Yamada and Peran (2018). Both studies layout high-level frameworks describing the issues data analytics governance should address. According to Avery and Cheek (2015) a data analytics governance framework should address issues such as human capital development and integration of data analytic activities in the organization. In our research, these issues are covered by the relational and structural mechanisms and the instantiations found in the case studies give more insight in the different ways how organizations can address these issues. Moreover, Yamada and Peran (2018) suggest a data analytics governance framework should contribute towards more alignment and development of analytics, which is addressed by the use of structural and process mechanisms in our framework.

Furthermore, this research provides examples on implementations of the four guiding principles (i.e. accountability, accessibility, community, and uniformity), which Avery and Cheek (2015) proposed for the development of an analytics framework. First, accountability on responsibilities is implemented by mechanisms concerning roles and responsibilities. Second accessibility to data analytics is implemented by

mechanisms concerning coordination and alignment. Third, community across the whole organization is implemented by the mechanism concerning organizational structure. Last, uniformity on policies is implemented by the mechanisms concerning monitoring and evaluation.

In addition, a deeper understanding of the maturity for data analytics governance is provided by an analysis on types of data analytics used (descriptive, predictive and prescriptive) and the number of mechanisms implemented. Case A is the organization that uses all three types of data analytics and also have the most mechanisms in place. The case organizations B and C are only using descriptive and predictive data analytics. They have in comparison lesser mechanisms in place than case A. This indicates that case A has a higher maturity of data analytics governance. For organizations, it is crucial to gain higher maturity in data analytics governance to build a competitive advantage (Dremel et al., 2017; Saltz & Shamshurin, 2016). Therefore, this study provides a first maturity perspective for data analytics governance.

From a practitioner's perspective, the results of the case studies can be considered valuable as it guides practitioners in defining an approach for data analytics governance. For example, practitioners could choose and customize a set of data analytics governance mechanisms (or instantiations thereof) from the framework, which they consider most appropriate for their organization.

There are also some limitations to take into account when using the results of this research. First of all, the limited amount of cases prevents validation of the framework and it leaves room to suppose that more mechanisms might be found in other organizations. Next, despite the fact that the use of a preliminary framework helped not being overwhelmed by the complexity of the situation, it could biased the collection and interpretation of data, and thereby limit identification of more mechanisms. Last, interview results were not used in subsequent interviews to check implementation of a mechanisms across cases. This limits validation if a specific mechanism implementation is also present at other case organizations.

As for future research, we plan to improve on this first framework for data analytics governance mechanisms. Therefore, future studies will use more in-depth case studies to discover new mechanisms and to understand the relations among mechanisms, because within some mechanisms organizations have to make trade-offs (e.g. central, hybrid, or decentralized organizational structure) and these are expected to influence the implementation of other mechanisms.

Another suggestion for future research is to develop a maturity model that links the use of certain governance mechanism to specific maturity levels. This idea is inspired by a case by case comparison of the data, which showed that some mechanisms are uniquely found in Case A, which is the organization that is most advanced in data analytics. Hence suggesting that only more advanced organizations are using these mechanisms but more research is needed to connect governance mechanisms to specific maturity levels.

Acknowledgements

We thank Paul van der Linden and Eelco Niens for their efforts in data collection and valuable contribution towards this research.

Developing a data analytics governance maturity model



The results presented in this chapter have been submitted to the Journal of Business and Information Systems Engineering as a full research paper.

Abstract

Leveraging data analytics is nowadays vital to creating value in organizations. At the same time, organizations find it challenging to reach the full potential of data analytics. To overcome these challenges, organizations need to govern their data analytics by providing clear structures, procedures, and guidelines. However, the realization of viable data analytics governance is neither easy nor straightforward. Accordingly, this study develops a data analytics governance maturity model to support organizations. A design science research is applied to design, demonstrate, and evaluate the maturity model in three different cycles. The development of the model yielded insights into the most important steps in the maturation of data analytics governance at the level of the organization. Furthermore, the results enable organizations to measure the maturity of their data analytics governance.

Keywords – *Data analytics; Governance; Maturity model; Roadmap*

6.1. Introduction

The rise of big data has prompted organizations to pay more attention to the use of data in business value creation (Grover et al., 2018). To extract value from data, organizations conduct different types of data analytics activities, which are often managed as projects. Within these projects, a group of persons with a mix of skills (e.g., mathematics, computer science, domain knowledge) complete a set of data analytics activities that convert data into insights or smart solutions (Davenport et al., 2012). However, conducting data analytics projects face several organizational challenges. One example is the need to align the incentives of management to those of data analytics practitioners. Managers prioritize securing a return on investments quickly, whereas practitioners prioritize accurate results (Yamada & Peran, 2018). Misalignment between management and analytics practitioners can result in failure. It can be overcome through clear procedures and guidelines on the conduct of data analytics activities. Incorporating appropriate governance is one means of achieving these ends (Rau, 2004; Weill & Ross, 2004). Among other things, governance concerns the allocation of decision rights and the establishment of procedures and policies (Gröger, 2018; Khatri & Brown, 2010; Rau, 2004). Therefore, data analytic activities should be governed to address organizational challenges (Gröger, 2018).

Previous research has formulated a framework for governing data analytics in organizations. That framework consists of a mix of mechanisms (Bajjens, Helms, & Velstra, 2020). On the first level, these mechanisms can be structural, relational, and process (He & Mahoney, 2006; Tallon et al., 2013; Zogaj & Bretschneider, 2014). On the second level of the framework, each of the mechanisms is decomposed into sub-mechanisms. Although the framework overviews the mechanisms, it does not identify the order in which they should be realized. Since sequencing is typically fraught with difficulty, organizations would benefit from further research into this topic. It would help them to avoid committing errors when governing data analytics.

Data analytics governance (DAG) is a comprehensive effort. Therefore, organizations must proceed incrementally. Maturity models are likely to prove helpful because they enable organizations to assess their current performance and to identify avenues for improvement (Bruin de et al., 2005). The basic concept of a maturity model comprises a set of areas in which the organization should progress. Progress is made along a predefined path to realize higher levels of maturity. High levels of maturity imply optimality (Smits, 2015), an important assumption that underlies the analysis. Organizations should strive to grow more mature in each area. The purpose of this research is to develop a maturity model to govern data analytics. The development of the maturity model will help to answer the following research question:

How should organizations progress in the governance of their data analytics?

The design science research (DSR) method is used to answer this question and to develop the maturity model (Hevner et al., 2004). The remainder of this paper is structured as follows. Section 6.2. presents the theoretical background, and Section 6.3. describes the DSR method. Thereafter, Section 6.4. presents the design, demonstration, and evaluation of the maturity model, as well as a refinement. Section 6.5. discusses the results of the three cycles and the critical path analysis. The conclusions are presented in Section 6.6., as are some suggestions for future research.

6.2. Theoretical background

This section provides a background to the concepts of data analytics, governance, and governance mechanisms. Thereafter, the section elaborates on existing maturity models that are applied in information systems research.

6.2.1. Data analytics governance

Data analytics aims to realize business value by using data to provide descriptive, predictive, and prescriptive insights. It draws on techniques from different disciplines, including software engineering, statistics, and machine learning (Delen & Demirkan, 2013; Lavallo et al., 2011). The data can be in the structured form that typifies conventional databases, with rows and columns, or unstructured (e.g., images or videos; (Larson & Chang, 2016). With the rise of new technologies, such as NoSQL databases and smart devices, unstructured data, also known as big data, became easier to collect and analyze (Chen et al., 2012), which fueled interest and investment (Watson, 2014). Among academics and practitioners, different terms for data analytics are used interchangeably. These terms include “data mining,” “business intelligence,” “advanced analytics,” “business analytics,” and “data science.” All refer to the practice of analyzing and exploring data to find new and interesting patterns, to improve decision-making, or to create smart solutions (Davenport, 2006).

The term “governance” refers to the procedures and practices by which a Board of Directors covers a variety of organizational issues to ensure that investments and activities are aligned with firm strategy (Rau, 2004). Governance can also focus on key organizational assets (Weill & Ross, 2004). For instance, IT governance is focused on how firms govern IT artifacts (Tallon et al., 2013), and data governance

focuses on governing data assets (Khatri & Brown, 2010). The increasing adoption of data analytics has intensified demand for a new focus to governance. While the concern of previous studies of governance was with IT artifacts (i.e., IT governance), or their content (i.e., information governance), little attention has been paid to the transformation of IT artifact content (De Haes & Van Grembergen, 2009; Fadler & Legner, 2021; Tallon et al., 2013). DAG concerns this very problem and aims to establish structures, policies, and controls to coordinate activities and to align interests so as to maximize the value of data analytics (Gröger, 2018; Yamada & Peran, 2018). The literature on governance asserts that it can be implemented through different mechanisms to attain this end (Baijens, Helms, & Velstra, 2020).

6.2.2. Data analytics governance mechanisms

Previous research has established a threefold typology of DAG mechanisms (Baijens, Helms, & Velstra, 2020). The framework, which is based on studies of IT and data, presented accounts for structural, process and relational governance mechanisms, as shown in Figure 6.1. (Alhassan et al., 2016; De Haes & Van Grembergen, 2009; He & Mahoney, 2006; Tallon et al., 2013; Zogaj & Bretschneider, 2014).

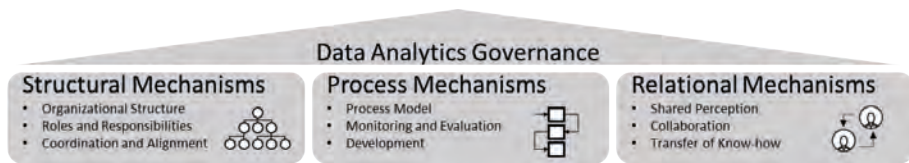


Figure 6.1. Framework for Data Analytics Governance (Baijens, Helms, & Velstra, 2020)

Structural governance mechanisms focus on organizing data analytics functions and related decision rights. They consist of three sub-mechanisms: organizational structure, roles and responsibilities, and coordination and alignment (Baijens, Helms, & Velstra, 2020). Firstly, organizational structure embeds data analytics within the organization so that the needs of different business units can be understood. The organizational structure can be decentralized, centralized, or hybrid (Grossman & Siegel, 2014). In a decentralized structure, the activities are spread across different units. In a centralized structure, all data analytics activities (e.g., decision-making problem prioritization) are situated in a single location. In a hybrid structure, coordination takes place in one location and other activities are diffused (Grossman & Siegel, 2014). Secondly, roles and responsibilities must be defined because successful data analytics demands new groups with diverse skills (Dremel et al., 2017; Grover et al., 2018). Those employed in different roles share responsibilities for creating leverage in the organizations. Those responsibilities may include managing

resources, monitoring data analytics activities, developing data analytics platforms, and designing data and information architecture (Schüritz et al., 2017). Finally, coordination and alignment are necessary. Data analytics activities are carried out across the organization (Espinosa & Armour, 2016). A dedicated committee could promote the business value of data analytics to ensure that projects obtain the necessary support and that viable candidate projects are prioritized (Dremel et al., 2018; Grossman, 2018; Grossman & Siegel, 2014; Kiron et al., 2011). The committee should consist of high-level managers to oversee data analytics and are responsible for the project portfolio (Dremel et al., 2017).

Process mechanisms are used to set up routines for the realization, monitoring, evaluation, and development of analytics processes. There are three process sub-mechanisms: process models, monitoring, and evaluation and development (Baijens, Helms, & Velstra, 2020). Process models are essential because they support the controlled and structured execution of projects. The best-known process model for data analytics is the CRISP-DM model (Mariscal et al., 2010). Another approach is to apply agile practices (Baijens, Helms, & Iren, 2020; do Nascimento & de Oliveira, 2012; Dremel et al., 2017; Schmidt & Sun, 2018). Although many organizations do not use a process model (neither from the literature nor a homegrown one), academics still agree that maintaining a well-defined, replicable process improves success rates (Saltz et al., 2018).

Monitoring and evaluating the efficiency and effectiveness of data analytics is necessary to improve consistency. Monitoring enables the organization to intervene when problems arise. Consequently, data analytics efforts produce the desired business results (Grossman & Siegel, 2014). Furthermore, evaluation allows organizations to measure return on investment and to connect data analytics efforts to business outcomes (Grover et al., 2018). Finally, a development roadmap should exist to ensure that data analytics activities develop in a direction that the organization deems desirable. As data analytics can have diverse goals, organizations should know how to reach them and how to improve their performance (Grossman, 2018).

Relational governance mechanisms should organize work in terms of relationships between people and groups. Three sub-mechanisms exist: shared perceptions, collaboration, and the transfer of knowledge and expertise (Baijens, Helms, & Velstra, 2020). As far as shared perceptions are concerned, organizations should prefer data over instinct and give themselves room to experiment and test (Abbasi et al., 2016; Berndtsson et al., 2018; Kiron et al., 2011). Support should be maintained even if initial results are disappointing. Perseverance requires the adoption of a solid organizational

attitude that is open to failure and alive to the uncertainty of the outcomes of data analytics. Organizations should provide sufficient autonomy to individuals so that they can judge their work (Barbour et al., 2018). Data analytics is perceived as multidisciplinary knowledge work, and it depends on collaboration between individuals with complementary skills to a considerable degree (Grover et al., 2018). The need to collaborate complicates management, and it makes communication more difficult. Therefore, organizations should promote communication between the individuals and groups that participate in data analytics activities (Barbour et al., 2018). Finally, it is crucial to transfer knowledge and expertise; organizations should ensure that they acquire and retain the right skills (Davenport et al., 2001). Therefore, they should foster the development of employee competencies. According to Dremel et al. (2018), organizations must implement central education programs to improve data analytics skills. Alternatively, they may hire external consultants.

6.2.3. Maturity models

Maturity models support organizations in the fluid implementation of governance. They provide firms with the ability to assess the status quo, to identify improvement measures, and to monitor progress (Bruin de et al., 2005). According to Röglinger et al. (2012), maturity is a state in which an organization can find itself and work continuously towards a dynamic goal.

In maturity models, this state is displayed by using different maturity levels (Becker et al., 2009). Maturity levels are sequential phases of organizational development. There are specific areas for improvement on each maturity level. Once improvement is accomplished, the company reaches the next level. Maturity models describe these levels and the corresponding maturity paths. Performance in different areas can be measured to ascertain the position of a company. If the areas are kept as generic as possible, the model can be applied to a large number of organizations (Röglinger et al., 2012).

Multiple maturity models for different practices have been developed in Information System research (Becker et al., 2009) more than a hundred maturity models have been developed to support IT management. They address a broad range of different application areas, comprising holistic assessments of IT management as well as appraisals of specific subareas (e. g. Business Process Management, Business Intelligence). Most of these models belong to one of two types, fixed-level models and focus-area models (Van Steenbergen et al., 2010). Most models in the literature are fixed-level models. Fixed-level models use five generic maturity levels, as shown in Table 6.1.. Each level represents several processes or practices that ought to be

implemented. This mode of presentation makes the models especially useful for benchmarking. However, there being no interdependencies between areas, fixed-level models are less useful for suggesting improvements. This makes it difficult to direct the development of an area.

Fixed-level maturity models take two different forms. Staged fixed-level models distinguish between a set number of generic levels of maturity. In continuous fixed-level models, areas are not attributed to a level. Instead, generic maturity levels are distinguished within each focus area. The capability maturity model (CMM) and its successor, the capability maturity model integration (CMMI), are the best-known fixed-level models (Becker et al., 2009; Röglinger et al., 2012). The five levels from these models are initial, repeatable, defined, managed, optimized, and widely adopted in other maturity models (Team, 2006). The levels are generic and can be applied to different areas to assess an organization.

Table 6.1. Fixed-level maturity model (based on Van Steenbergen et al. (2007))

	Level 1	Level 2	Level 3	Level 4	Level 5
Area 1	A	B	C	D	E
Area 2	A	B	C	D	E
Area 3	A	B	C	D	E
Area 4	A	B	C	D	E

Focus-area models are relatively new. They are premised on the incremental improvement of a collection of items (Van Steenbergen et al., 2010; Van Steenbergen et al., 2007). This type of model does not use generic levels. Maturity items are attached to each focus area instead, as shown in Table 6.2.. In order to ensure that the model represents the overall level of maturity of an organization, maturity is expressed as a combination of items. In each focus area, there are steps, in the form of maturity items, which build up to maturity. The maturity items are specific to certain focus areas. This is the main difference from fixed-level models, where generic levels are used for different areas. As noted earlier, fixed-level models provide a simple overview of maturity levels, but no detailed directions for improvement. This gives focus-area models more freedom to use different levels in different areas, which is important because there can be large differences between specific implementations.

Table 6.2. Focus-area maturity model (based on Van Steenberghe et al. (2007))

	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7
Focus area A	AA				AB		
Focus area B		BA		BB			
Focus area C	CA		CB			CC	
Focus area D				DA			DB

6.3. Methodology

The research methodology chosen for this study was Design Science Research (DSR). It is an effective problem-solving methodology for research, by using evaluation, and communication, and rigor practices (Hevner et al., 2004). Furthermore, its aim is to yield artefacts that create design knowledge (Wieringa, 2014). This goal is accomplished by combining existing and new design knowledge in multiple “build” and “evaluate” activities in order to create innovative solutions to a problem (vom Brocke et al., 2020).

In this research, DSR is applied to design, demonstrate, and evaluate a maturity model in a real-life setting and to understand how organizations govern data analytics. Applying DSR results in prescriptive knowledge about DAG. Their application also shows how the maturity model can be used, how DAG develops in organizations (Drechsler & Hevner, 2018; vom Brocke et al., 2020), and how governance practices mature (Hevner et al., 2004).

The DSR methodology contains specific steps for the development and evaluation of an artefact (Peppers et al., 2007). The steps were followed during this study and consisted of identifying the problem, defining the objective of the solution, as well as design, demonstration and evaluation. These steps are common in design science and have been applied in other studies (Ahangama & Poo, 2015a; Kloör et al., 2018; Volk et al., 2017). Following these steps helped to create a rigorous maturity model that is relevant to the conduct of data analytics in organizations. The remainder of this section overviews them.

For the steps “identify the problem” and “define the objective of the solution,” this research contributes to reducing the difficulties that organizations face in governing data analytics by creating a maturity model (Bruin de et al., 2005; Gröger, 2018). As noted in the introduction, this is the problem and the solution for the study. The steps “design,” “demonstration,” and “evaluation” are conducted in multiple cycles

to develop the maturity model. To create an effective maturity model, it is necessary to identify different areas, items, and logical relationship between those items (Pöppelbuß & Röglinger, 2011; Van Steenberg et al., 2010; Van Steenberg et al., 2007). To identify these areas, items and relationships to create the model for this study, three cycles of design, demonstration, and evaluation were executed. In the first cycle, the aim was to identify areas and individual maturity items. In the second, the aim was to explore the interdependencies between the items. In the third cycle, the model was applied to real-life cases to test its validity. Figure 6.2. overviews the three cycles and the DSR steps.

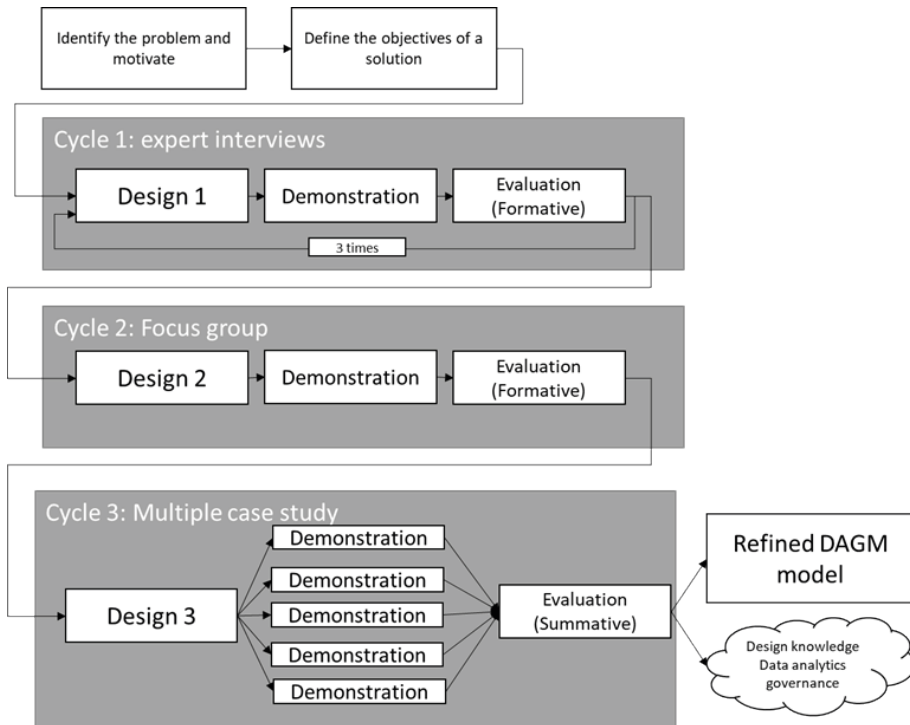


Figure 6.2. Design Science Research Process

Evaluation is an important step in the DSR approach because it determines whether the artefact is a solution to the identified problem. It is at this stage that new knowledge is created, which leads to the adoption of the artefact, depending on the findings. Therefore, it is important to use a sound approach for the evaluation stage in DSR. Venable et al. (2016) developed several strategies for evaluation. The technical risk efficacy strategy was used in this research (see Figure 6.3.). Its application resulted in the three cycles that are displayed in Figure 6.2.. The first

cycle represents “artificial evaluation” in the technical risk efficacy strategy. This first evaluation does not involve real users, that is, it is artificial, because the applied model in the first cycle is based on a rough design. In this study, accordingly, the initial evaluation was restricted to an artificial setting, with expert interviews and focus groups as its bases. Thereafter, a more summative case-study evaluation was used to determine the validity with which the model measures DAG outcomes in a naturalistic environment (Prat et al., 2014). The validity of the model was adequate when the results are recognized in real-life organizations. The next section, discusses how each of the cycles was conducted and how the artefact and knowledge of it evolved during these steps.

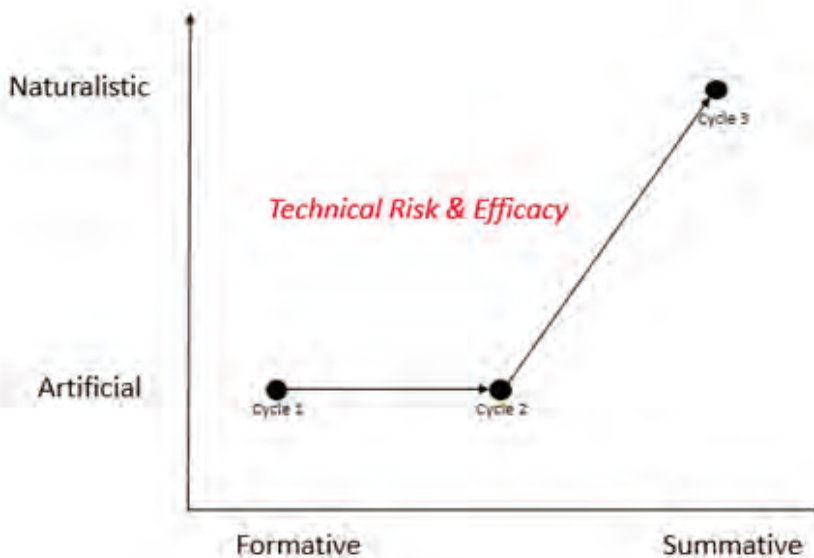


Figure 6.3. Evaluation Strategy (Based on (Venable et al., 2016))

6.4. Design cycles: results

This section presents the results of the design, demonstration, and evaluation in the three cycles. Then, the refined data analytics governance model (DAGM) is presented.

6.4.1. Results: Cycle 1

Previous research has already identified the mechanisms of the DAG framework (Baijens, Helms, & Velstra, 2020). Therefore, the first design in Cycle 1 used the mechanisms of the DAG framework as an input for a maturity model. A choice had to be made between fixed-level and focus-area models, as discussed in Chapter 2. When designing maturity models, the balance between simplification and verisimilitude is vital. Miscalibration casts the usability and applicability of the maturity model into doubt (Bruin de et al., 2005). To ensure that the maturity model was balanced well, a focus-area model was chosen: the DAG framework indicates that different hierarchical levels exist within its sub-mechanisms (Baijens, Helms, & Velstra, 2020). Using a fixed-level maturity model causes generic levels to be applied across all mechanisms. Consequently, information is lost. Therefore, the focus-area maturity model is the most appropriate (Table 6.2.) for this study. Such models make it possible to distinguish between more than five levels of maturity. As a result, the gaps between the levels are small, and more guidance is available on incremental improvements (Jansen, 2020; Smits & Hillegersber van, 2017; Spruit & Röling, 2014; Van Steenbergen et al., 2010; Van Steenbergen et al., 2007). This feature of focus-area models is especially relevant in the new area of DAG, where organizations have no clear perspective of potential ways to improve. A focus-area model supports an incremental approach to maturity in specific areas. Therefore, the DAG framework was combined with the focus-area maturity model to create a first design of the Data Analytics Governance Maturity (DAGM) model, as shown in Table 6.3.. The three mechanisms (the structural, the process and the relational) from the DAG framework were used as focus areas. The illustrations of the mechanisms identified by Baijens, Helms and Velstra (2020) were used in the selection of the maturity items for the focus areas. This design was not a functioning focus-area maturity model because the focus areas were independent.

Table 6.3. First Design of DAGM Model

Focus area	Maturity Item
A. Process	A.0. The organization conducts data analytics projects in an ad-hoc fashion.
	A.1. The organization has a formalized process model and/or project methodology for guiding data analytics projects.
	A.2. The organization monitors the execution of projects by measuring the progress of the data analytics projects.
	A.3. The organization evaluates the outcomes of data analytics projects against business objectives.
	A.4. The organization has a formalized demand management process for the allocation of resources to data analytics projects.
	A.5. The organization uses feedback to improve data analytics processes and/or project methodology continuously.
	A.6. The organization translates its data analytics strategy into a concrete plan for action (i.e., a roadmap).
B. Structure	A.7. The organization evaluates its data analytics strategy at regular intervals.
	B.0. The organization takes decisions about data analytics projects in an ad-hoc fashion.
	B.1. The organization has defined specific roles that are involved in the execution and management of data analytics projects.
	B.2. The organization has defined responsibilities for the execution and management of data analytics projects.
	B.3. Data analytics projects are conducted in close collaboration with the business.
C. Relation	B.4. The organization coordinates its data analytics project portfolio through a governing body (e.g., a committee or a project group).
	C.0. The organization takes decisions about data analytics projects in an ad-hoc fashion.
	C.1. The organization has defined its data analytics strategy.
	C.2. There is support for the data analytics strategy within the organization.
	C.3. The organization initiates frequent meetings to support collaboration and discussions for the alignment of data analytics (e.g., stand-up meetings).
	C.4. The organization offers educational programs to train employees in the use of data analytics.
C.5. The organization supports different modes of data analytics knowledge transfer (e.g., internal/external conferences, job rotation).	



The first design was demonstrated in three consecutive expert interviews. The interviews were conducted online due to the Covid-19 pandemic. The participants worked in data analytics environments and had master's degrees in data science management. The interviews were guided by an interview protocol that was based on the maturity items from the first design. They lasted for one hour on average and were recorded for subsequent analysis. The expert-interview demonstration circled on understanding the focus areas and the set of maturity items within the focus group as well as on identifying the correct hierarchy of the maturity items. The sequential conduct of the interviews was advantageous because the output from one interview could be used as an input for a new design and presented in the subsequent interview. This resulted in the refinement of maturity items, multiple suggestions for adjustment, and the addition of new maturity items, as shown in Table 6.4.. The evaluation in the first cycle had a formative purpose and took place in an artificial setting. Expert insights were used to improve the artefact.

Table 6.4. Evaluation Results: DAGM Model (First Design)

<p>Adjustments to model:</p> <ul style="list-style-type: none"> • Use of eight sub-mechanisms as focus areas instead of three. • Removal of formalized demand management process for the allocation of resources to data analytics projects. • Removal of ad-hoc levels. • Merger of maturity items "roles," "responsibilities," and "organization structure."
<p>Added maturity items:</p> <ul style="list-style-type: none"> • The organization must use process models that suit the situation best. • The organization must have a recognizable place for data analytics. • The whole organization uses a formalized process model and/or project methodology for data analytics projects. • The organization monitors the end results of data analytics projects to provide maintenance. • The organization uses external educational programs to train employees in new data analytics techniques. • The organization offers educational programs (internal/external) to broaden data analytics knowledge. • The organization overviews its data analytics project portfolio regularly.

6.4.2. Results: Cycle 2

In Cycle 2, the design integrated the focus area items and the improved set of maturity items. Their hierarchy was adjusted to reflect inputs from the previous cycle, as shown in Table 6.5.a. and 6.5.b.

Table 6.5.a. Second Design of DAGM model

Mechanism	Focus area	Maturity items
A. Process	A.A. Process model	A.A.1. Has the organization adopted a formalized process model and/or project methodology for data analytics projects?
		A.A.2. Does the whole organization use a formalized process model and/or project methodology for data analytics projects?
		A.A.3. Does the organization select its process model and/or project methodology depending on the project that it intends to execute?
		A.A.4. Does the organization use feedback to improve its data analytics processes and/or project methodology continuously?
	A.B. Monitoring and evaluation	A.B.1. Does the organization monitor the execution of projects by measuring their progress against their objectives?
		A.B.2. Does the organization evaluate the outcomes of data analytics projects against business objectives?
		A.B.3. Does the organization monitor the final results of data analytics projects continuously?
	A.C. Development	A.C.1. Does the organization hold meetings to reflect on the data analytics strategy?
		A.C.2. Does the organization have a concrete plan of action (i.e., a roadmap) to achieve the goal of its data analytics strategy?
		A.C.3. Does the organization evaluate its actions against the expectations recorded in the roadmap regularly?
B. Structure	B.A. Organization structure, roles, and responsibilities	B.A.1. Has the position of data analytics within the organization been defined?
		B.A.2. Has the organization defined roles and responsibilities for the execution and management of the documented data analytics projects?
		B.A.3. Has the organization defined roles and responsibilities for the execution and management of data analytics projects that are implemented?
		B.A.4. Does the organization regularly review/update the roles and responsibilities for the execution and management of data analytics projects?
	B.B. Coordination and alignment	B.B.1. Does the organization overview its data analytics project portfolio?
		B.B.2. Does the organization coordinate data analytics project portfolio through a governing body (e.g., a committee or a project group)?

Table 6.5.b. Second Design of DAGM model

Mechanism	Focus area	Maturity items
C. Relation	C.A. Shared perceptions	C.A.1. Is the management of the organization aware of the opportunities of data analytics without forcing the organization to exploit them?
		C.A.2. Are the managers of the organization convinced of the value of data analysis, and do they see it as crucial?
		C.A.3. Is data analytics used in decision-making throughout the organization?
	C.B. Collaboration	C.B.1. Does the organization provide its employees with opportunities (physical or digital) to communicate and collaborate easily?
		C.B.2. Does the organization initiate regular meetings to support collaboration and discussions of the alignment of data analytics (e.g., stand-up meetings)?
		C.B.3. Are data analytics projects in co-creation with the business?
	C.C. Transfer of know-how	C.C.1. Does the organization offer educational programs to train employees in new techniques and to broaden their knowledge of data analytics?
		C.C.2. Does the organization use educational programs to train employees in new techniques and to broaden their knowledge of data analytics?
		C.C.3. Does the organization support different modes of data analytics knowledge transfer (e.g., internal/external conferences, job rotations)?

The demonstration was conducted with a focus group. The group consisted of the three interviewees from the first cycle. The meeting was convened online due to the Covid-19 pandemic. An online whiteboard was used. The focus group meeting lasted two hours, and the video was recorded with Microsoft Teams for subsequent analysis. The discussion generated a degree of consensus about the design of the DAGM model. It centered on the dependencies between maturity items from different focus areas. The discussion resulted in the design of a DAGM that accounts for those dependencies, as shown in Table 6.6.. For example, the maturity item “Is the management of the organization aware of the opportunities of data analytics without forcing the organization to exploit them?” (C.A.1) should be considered before the item “Has the position of data analytics within the organization been defined?” (B.A.1). The purpose of the second-cycle evaluation was formative. Therefore, the focus group yielded data from an artificial setting.

Table 6.6. Evaluation Results: DAGM Model (Second Design)

Maternity item	Input
A.A.1. Has the organization adopted a formalized process model and/or project methodology for data analytics projects?	C.B.2.
A.A.2. Does the whole organization use a formalized process model and/or project methodology for data analytics projects?	none
A.A.3. Does the organization select its process model and/or project methodology depending on the project that it intends to execute?	A.B.3.
A.A.4. Does the organization use feedback to improve its data analytics processes and/or project methodology continuously?	none
A.B.1. Does the organization monitor the execution of projects by measuring their progress against their objectives?	A.B.2.
A.B.2. Does the organization evaluate the outcomes of data analytics projects against business objectives?	A.C.1., C.A.2.
A.B.3. Does the organization monitor the final results of data analytics projects continuously?	A.A.4.
A.C.1. Does the organization hold meetings to reflect on the data analytics strategy?	A.C.2.
A.C.2. Does the organization have a concrete plan of action (i.e., a roadmap) to achieve the goal of its data analytics strategy?	A.A.3., A.C.3., C.A.3.
A.C.3. Does the organization evaluate its actions against the expectations recorded in the roadmap regularly?	none
B.A.1. Has the position of data analytics within the organization been defined?	A.A.1., B.A.2., B.B.1., C.B.1., C.C.1.
B.A.2. Has the organization defined roles and responsibilities for the execution and management of the documented data analytics projects?	B.A.3.
B.A.3. Has the organization defined roles and responsibilities for the execution and management of data analytics projects that are implemented?	A.A.2., A.B.1., B.B.2.
B.A.4. Does the organization regularly review/update the roles and responsibilities for the execution and management of data analytics projects?	none
B.B.1. Does the organization overview its data analytics project portfolio?	none
B.B.2. Does the organization coordinate data analytics project portfolio through a governing body (e.g., a committee or a project group)?	none
C.A.1. Is the management of the organization aware of the opportunities of data analytics without forcing the organization to exploit them?	B.A.1.
C.A.2. Are the managers of the organization convinced of the value of data analysis, and do they see it as crucial?	B.A.4., C.B.3., C.C.2.
C.A.3. Is data analytics used in decision-making throughout the organization?	A.B.3.
C.B.1. Does the organization provide its employees with opportunities (physical or digital) to communicate and collaborate easily?	C.B.2.
C.B.2. Does the organization initiate regular meetings to support collaboration and discussions of the alignment of data analytics (e.g., stand-up meetings)?	B.B.2.
C.B.3. Are data analytics projects in co-creation with the business?	none
C.C.1. Does the organization offer educational programs to train employees in new techniques and to broaden their knowledge of data analytics?	none
C.C.2. Does the organization use educational programs to train employees in new techniques and to broaden their knowledge of data analytics?	none
C.C.3. Does the organization support different modes of data analytics knowledge transfer (e.g., internal/external conferences, job rotations)?	none

6.4.3. Results: Cycle 3

The third cycle entailed the allocation of the maturity items to the focus areas that were identified in the previous cycle. This revealed the dependencies between the maturity items of the focus areas in sequential order, as shown in Figure 6.4.. The aim of this cycle is to evaluate the maturity model and to determine how validly it measures DAG. Therefore, the design was demonstrated in a multiple case study, with interviews at five organizations (A, B, C, D and E). The organizations were selected through convenience sampling. They were active contacts of the research center where the research was conducted. Organization A, Organization B, and Organization E were public organizations, and Organization C and Organization D were commercial organizations, as shown in Table 6.7..

Mechanism	Focus area	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	Level 10	Level 11
A. Processes area	A.A. Process model			A.A.1.		A.A.2.				A.A.3.		A.A.4.
	A.B. Monitor & evaluation					A.B.1.	A.B.2.				A.B.3.	
	A.C. Development							A.C.1.	A.C.2.	A.C.3.		
B. Structural area	B.A. Organization structure, roles and responsibilities		B.A.1.	B.A.2.	B.A.3.				B.A.4.			
	B.B. Coordination and alignment			B.B.1.		B.B.2.						
C. Relational area	C.A. Shared perceptions	C.A.1.						C.A.2.		C.A.3.		
	C.B. Collaboration			C.B.1.	C.B.2.				C.B.3.			
	C.C. Transfer of know how			C.C.1.					C.C.2.			C.C.3.

Figure 6.4. Third Design of DAGM model

All of the organizations were experienced in the execution of data analytics projects and had enjoyed considerable success. Interviews were conducted to determine how each organization would score on the DAGM model. The interviewees were responsible for the data analytics project portfolio at the organizations and were therefore capable of scoring the maturity of their employers accurately. Some interviewees were unsure if they could answer all questions with sufficient accuracy. Therefore, additional interviews were conducted at Organization A and at Organization B. The interviews were conducted online due to the Covid-19 pandemic. A DAGM model scoresheet was used.



Table 6.7. Case descriptions

Organization	Industry	Employees	Position of interviewee(s)
A	Public	2,900	- Innovation manager - Lead intelligence
B	Public	3,000	- Skill lead (data science and analytics) - Product Owner (actionable insights)
C	Transport	1,600	- Process and product manager
D	Financial	1,500	- Chief analytics officer
E	Public	2,000	- Head process developer

The interviewees were informed of the purpose of the interview in advance. During its course, they were briefed on the maturity model. The scoresheet was completed by answering the researcher's questions. The participants could answer if a maturity item was present in their organization by saying "yes," "no," or "partially." This demonstration yielded the scoresheet shown in Table 6.8.. The green boxes indicate that the corresponding maturity item was present at an organization. The red boxes indicate absence. The yellow boxes indicate that the implementation of the corresponding item was ongoing.

After the completion of the scoresheet, the interviewees were given time to reflect on the scores in an open discussion and to determine whether the suggestions of the model were useful for the organization. Furthermore, the interviewees could suggest improvements to the design. These improvements included gaps and ambiguities in the model. The focus of the case study was to ascertain the validity of the DAGM model for measuring DAG maturity.

Table 6.8. Evaluation Results: DAGM Model (Third Design)

Case	Sub area	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Level 7	Level 8	Level 9	Level 10	Level 11
A	A.A.			A.A.1.		A.A.2.				A.A.3.		A.A.4.
	A.B.					A.B.1.	A.B.2.				A.B.3.	
	A.C.							A.C.1.	A.C.2.	A.C.3.		
	B.A.		B.A.1.	B.A.2.	B.A.3.				B.A.4.			
	B.B.			B.B.1.		B.B.2.						
	C.A.	C.A.1.						C.A.2.		C.A.3.		
	C.B.			C.B.1.	C.B.2.				C.B.3.			
	C.C.			C.C.2					C.C.2.			C.C.3.
B	A.A.			A.A.1.		A.A.2.				A.A.3.		A.A.4.
	A.B.					A.B.1.	A.B.2.				A.B.3.	
	A.C.							A.C.1.	A.C.2.	A.C.3.		
	B.A.		B.A.1.	B.A.2.	B.A.3.				B.A.4.			
	B.B.			B.B.1.		B.B.2.						
	C.A.	C.A.1.						C.A.2.		C.A.3.		
	C.B.			C.B.1.	C.B.2.				C.B.3.			
	C.C.			C.C.4					C.C.2.			C.C.3.
C	A.A.			A.A.1.		A.A.2.				A.A.3.		A.A.4.
	A.B.					A.B.1.	A.B.2.				A.B.3.	
	A.C.							A.C.1.	A.C.2.	A.C.3.		
	B.A.		B.A.1.	B.A.2.	B.A.3.				B.A.4.			
	B.B.			B.B.1.		B.B.2.						
	C.A.	C.A.1.						C.A.2.		C.A.3.		
	C.B.			C.B.1.	C.B.2.				C.B.3.			
	C.C.			C.C.5					C.C.2.			C.C.3.
D	A.A.			A.A.1.		A.A.2.				A.A.3.		A.A.4.
	A.B.					A.B.1.	A.B.2.				A.B.3.	
	A.C.							A.C.1.	A.C.2.	A.C.3.		
	B.A.		B.A.1.	B.A.2.	B.A.3.				B.A.4.			
	B.B.			B.B.1.		B.B.2.						
	C.A.	C.A.1.						C.A.2.		C.A.3.		
	C.B.			C.B.1.	C.B.2.				C.B.3.			
	C.C.			C.C.6					C.C.2.			C.C.3.
E	A.A.			A.A.1.		A.A.2.				A.A.3.		A.A.4.
	A.B.					A.B.1.	A.B.2.				A.B.3.	
	A.C.							A.C.1.	A.C.2.	A.C.3.		
	B.A.		B.A.1.	B.A.2.	B.A.3.				B.A.4.			
	B.B.			B.B.1.		B.B.2.						
	C.A.	C.A.1.						C.A.2.		C.A.3.		
	C.B.			C.B.1.	C.B.2.				C.B.3.			
	C.C.			C.C.7					C.C.2.			C.C.3.

As discussed previously, validity is the main criterion for evaluating the DAGM model because there are no DAG maturity models in existence. The validity assessment was summative. First, the from the DAGM model were discussed with the interviewee to gauge their accuracy. Second, organizations where the sequence of maturity items in a given focus area differed from that suggested by the DAGM model could indicate that the model was invalid. These instances were discussed with the organizations. The yellow sections were counted to produce maturity scores. All successive maturity items need to be implemented for a specific level to be completed. Advice reflects this concept. For example, Organization A had attained Level 6 because all relevant maturity items were present. For the organization to mature in DAG, it would need to implement the maturity item "Are the managers of the organization convinced of the value of data analysis, and do they see it as crucial?" (C.A.2.). These results, which are shown in Table 6.9., were shared with the organizations, as were the suggestions.

Reviewing this score led to the following perspectives about the validity of the model: Organization D and Organization E indicated that the advice was accurate, and ideas from the advice was something they had been working on but had thus far failed to achieve. Organization A also indicated that the advice was accurate, although they stated that it would be difficult for a public organization such as theirs to act on it. Organization B and Organization C found the advice reasonable. However, they did indicate that parts of the model were difficult to measure and met these conditions. Therefore, they suggested improving the model. Their suggestion was adopted in the formative evaluation.

Table 6.9. DAGM model case organization score's

Case	Level	Advice
A	6	- Are the managers of the organization convinced of the value of data analysis, and do they see it as crucial?(C.A.2.)
B	7	- Does the organization evaluate its actions against the expectations recorded in the roadmap regularly? (A.C.3.) - Is data analytics used in decision-making throughout the organization? (C.A.3.)
C	3	- Has the organization defined roles and responsibilities for the execution and management of data analytics projects that are implemented? (B.A.3.)
D	7	- Does the organization have a concrete plan of action (i.e., a roadmap) to achieve the goal of its data analytics strategy? (A.C.2.)
E	7	- Does the organization regularly review/update the roles and responsibilities for the execution and management of data analytics projects? (B.A.4.)

The second approach to checking the validity of the DAGM involves determining whether the order of the maturity items corresponds to reality. A valid maturity model would only include consecutive maturity items. Therefore, in Table 6.8., red or yellow cells would not be followed by green ones. The examination of the order of the items revealed five instances that deviated from the order of the DAGM model across three different focus areas, namely “process model” (Organization C), “transfer of know-how” (Organization B) and “monitoring and evaluation” (Organization B, Organization C, and Organization E). These results were discussed with interviewees from the organizations. The discussion helped identify counter indications to assuming dependencies between maturity items.

The focus area “process model” assumes that the organization should adopt a formalized process model or project methodology (A.A.1.) before using one (A.A.2.). Then, the organization can choose a process model or a project methodology that suits the project (A.A.3.). Last, the organization uses feedback to improve its data analytics processes or its project methodology continuously (A.A.4.). Organization C lacks a formalized methodology (A.A.1. and A.A.2) but can choose between different methodologies for particular project types (A.A.3). Its representatives pointed out that some methods were formalized and that others were not, and their approaches to small and large projects differed. However, formalizing a process model or a methodology in general terms would contribute to monitoring progress. Organization C agreed that a more consistent working method with a formal methodology would be useful, but they had not yet incorporated this approach into their work.

In the focus area “transfer of know-how,” the DAGM model suggests that organizations should implement educational programs to broaden their knowledge base (C.C.3) before using different modes of knowledge transfer (C.C.4). However, the opposite appears to have been true of Organization B. The interviewee indicated that it is much easier to realize knowledge sharing in a small community than across an organization. However, changing the corresponding sequence in would conflict with the observations for Organization C. Therefore, the definition in the maturity item should be adjusted so as not to focus on the whole organization.

In the “monitoring and evaluation” focus area, there was a misalignment between the order of the design and three of the cases. According to the design, organizations should begin by monitoring projects (A.B.1.). Then, they should evaluate them (A.B.2.) and finally, monitor the results of a project for maintenance (A.B.3). However, in Organization B, maintenance was monitored, but the projects were not evaluated sufficiently. Organization B indicated that their approach was specific and that the misalignment should not be taken to imply that the model was invalid.

In Organization C and Organization E, evaluations would precede monitoring. At Organization C, this approach would only be adopted for significant projects, which is why the organization only indicated partial compliance with the item. The representative of Organization E mentioned that they had encountered problems with this approach in the past, when projects ran for so long that they became unfeasible. The interviewee thought that it was acceptable to evaluate without monitoring. However, monitoring is necessary to acquire a broader perspective.

After the presentation of the model scores, the interviewees reflected on the model to discuss omissions. Some of the resultant suggestions were adopted in the refined DAGM model. Organization B suggested that it might be helpful to check the reliability of the analytical results. Thus, it is important not only to ensure that the results of a data analytics project are evaluated against business objectives but also to evaluate the reliability of the data, that is, one must determine the origin of the data and the process by which it was created. The purpose of this exercise is to prevent the use of invalid data analytics results. This suggestion pertains to the focus area “monitoring and evaluation.” The Organization D representative suggested improving item C.C.3, “Does the organization support different modes of data analytics knowledge transfer (e.g., internal/external conferences, job rotations)?”. According to them, the maturity item is unambitious. This is the last level, and the organization should be proactive instead of merely “supporting” knowledge transfer. The suggestion was incorporated into the refined DAGM model.

6.4.4. Refined DAGM model

The suggestions were used to refine the DAGM model. The refined DAGM model has 11 levels. It can function as a roadmap that practitioners can use to implement DAG and to direct its realization. The levels are presented in Table 6.10.. Each level consists of items that can be completed in parallel.

Table 6.10. Refined Data Analytics Governance Maturity Model

Level 1	The organization should create managerial awareness of data analytics opportunities (C.A.1).
Level 2	The position of the data analytics function in the organization should be defined (B.A.1).
Level 3	The organization should focus on the formalization of a process model or a project methodology (A.A.1), document roles and responsibilities (B.A.2), and overview data analytics projects (B.B.1). Furthermore, it should provide accessible avenues of accessible communication (C.B.1) and educational programs to train employees in data analytics (C.C.1).
Level 4	The organization should implement roles and responsibilities (B.A.3) and initiate regular meetings to support collaboration and discussions of the alignment of data analytics (C.B.2).
Level 5	The organization should focus on using a formalized process model or project methodology (A.A.2) and monitor the execution of data analytics projects (A.B.1). In addition, it should coordinate the portfolio of data analytics projects through a governing body (B.B.2).
Level 6	The organization should evaluate the outcomes of data analytics projects against business objectives and reliability (A.B.2).
Level 7	The organization should initiate meetings to reflect on the data analytics strategy (A.C.1), and management should see data analytics as crucial (C.A.2).
Level 8	The organization should have a roadmap for the implementation of its data analytics strategy. (A.C.2) It should review/update the roles and responsibilities of staff members who participate in data analytics projects regularly (B.A.4), and it should conduct data analytics projects in co-creation with the business (C.B.3). Moreover, it should use educational programs to teach employees about data analytics (C.C.2).
Level 9	The organization should choose a process model or a project methodology depending on the project that they plan to execute (A.A.3) and evaluate its actions against the expectations recorded in the roadmap regularly (A.C.3). Moreover, data analytics should be used throughout the organization (C.A.3).
Level 10	The organization should monitor the results of data analytics projects and provide maintenance to ensure alignment with business objectives (A.B.3).
Level 11	The organization should improve its data analytics processes or its project methodology continuously (A.A.4). Furthermore, it should employ different modes of data analytics knowledge transfer (C.C.3).

6.5. Discussion

The primary objective of this study was to improve the understanding of the implementation of DAG by developing a DAGM model. This section will begin with a discussion of the validity of the model. The discussion is based on the results of the three rounds of evaluation. A critical path analysis is applied to the dependencies between the items in the final version of the DAGM model to identify the most essential steps in the development of DAG in an organization.

6.5.1. Validity of the DAGM model

The initial DAGM model was designed on the basis of the literature on DAG mechanisms. In the first cycle, the model was evaluated through three expert interviews. The experts indicated how the validity of the model could be improved and how it could be aligned to practice. Validity was assessed further in the second cycle with a focus group of three experts. The focus group provided practical insights about the sequence of the maturity items, which led to improvements. The evaluation of the model in the third cycle showed its validity for measuring the maturity of DAG. Three organizations (A, D, and E) stated that they could recognize that their organizations belonged to particular levels within the model. The suggestions for each level were also received well. The interviewees indicated that the suggestions sounded familiar and that it was an aspect they were processing. Two organizations (B and C) indicated that they could recognize that their organizations was at a certain level, but they mentioned that using the model was sometimes difficult. Accordingly, they suggested some improvements.

In addition, examining the order in which the organizations completed maturity items with that assumed in the DAGM model provided additional insights about the validity of the model. Five deviations were identified. However, all five could be explained and neither reflected shortcomings of the DAGM model. Instead, they were associated with the specific characteristics of the organizations under observation.

One limitation of the validity assessment presented here is that the companies were not followed over a long period. A more comprehensive study would yield more concrete indications of the verisimilitude of the maturation scheme that the model adopts. Furthermore, while the demonstration and the evaluation enabled the dependency of the maturity items to be measured in each focus area, dependency across focus areas was too complex to fall within the scope of the study.

6.5.2. Maturity stages of data analytics governance

The analysis of the dependencies (Figure 6.4.) reveals that some items seem to be on the critical path of DAGM progress. An item is said to be on this critical path if other items cannot commence before that item is completed. Items on the critical path need to be safeguarded because delaying these items cause delays in the entire chain (Willis, 1985). In terms of the DAGM model presented here, delays would postpone the attainment of the highest maturity level. Therefore, the critical path indicates the items that ought to be monitored closely if delays are to be avoided and progress ensured. The following paragraphs discuss the items that were identified as parts of the critical path of the DAGM model. Of those, there are six, namely creating awareness, structuring, measuring, long-term planning, adapting, and continuing. Each of these stages is shown as a grey area in Figure 6.5..

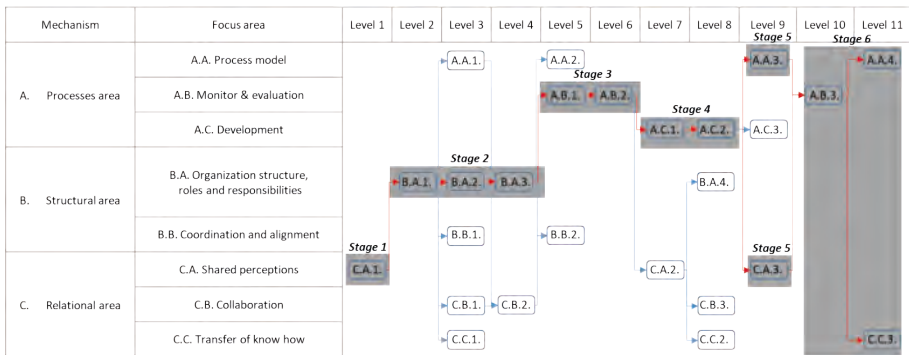


Figure 6.5. Critical Path Analysis (DAGM Model)

First, the awareness creation stage consists of one item, which is called “management awareness” (C.A.1., Figure 6.5.). This item indicates that the first step towards DAGM is raising awareness of DAG among managers. The importance of management awareness is also addressed in other studies of the conditions for the success of data analytics (Chen et al., 2017; Gao et al., 2015; Vosloo & Naidoo, 2019). Consequently, it is unsurprising that awareness is the starting point of establishing successful DAG.

The second stage concerns structuring. The organization must examine the locations where data analytics activities are formally performed and the roles and responsibilities of individual actors (B.A.1., B.A.2., B.A.3., Figure 6.5.). Because management awareness has already been generated, it is easier to implement a

structure and to maintain it through top-down control. The relevant items all pertain to the structural mechanisms of governance. Other studies identify them as crucial for the governance of analytics (Fadler & Legner, 2021; Schüritz et al., 2017).

In the third stage, the organization must engage in measurement by monitoring and ultimately evaluating its activities to ensure that the results contribute to the business objective (A.B.1., A.B.2., Figure 6.5.). The roles created in the previous phase facilitate this endeavor because monitoring and evaluation responsibilities can be divided. The ability to monitor and evaluate allows an organization to respond quickly to changing circumstances (Grossman, 2018; Grover et al., 2018).

In the fourth stage, the organization must engage in long-term planning by deciding whether its strategy is being successful and planning further progress (A.C.1., A.C.2., Figure 6.5.). This also makes sense because the results of the evaluation can serve as inputs for determining whether the strategy is being executed adequately.

In the fifth stage, the organization focuses on becoming more flexible. Having the long-term planning and measurement in place for the previous stages contributes to this and the organization can focus on increasing the number of employees who use data analytics and on adapting its way of working to its goal. Consequently, the organization may tailor its approach to individual projects (A.A.3., Figure 6.5.; (Baijens, Helms, & Kusters, 2020)), and a data-driven culture may spread among its workforce (C.A.3., Figure 6.5. Critical Path Analysis (DAGM Model)5.; (Berndtsson et al., 2018; Grover et al., 2018; Vidgen et al., 2017)).

In the sixth, and final, stage, the organization focuses on continuing its activities. The stage comprises Level 10 and Level 11. First, current results should be monitored and maintained (A.B.3., Figure 6.5.). Next, the organization should be arranged in a way that enables it to improve the process and to train its staff continuously (A.A.4., C.C.3., Figure 6.5. Critical Path Analysis (DAGM Model)5.).

6.6. Conclusion

This paper presented a DAGM model which is informed by a DSR approach. The development of the maturity model provided a deeper understanding of the governance of data analytics and its trajectory to maturity. The results reflect three cycles of design, demonstration, and evaluation. The final cycle involved summative evaluation and showed that the model was perceived as a valid measurement of DAG maturity. This evaluation added to the knowledge of DAG and culminated in a roadmap for the governance of data analytics. Furthermore, a critical path was identified. It consists of creating awareness, structuring, measuring, long-term planning, adapting, and continuing.

The research also has scientific value. The model yielded insights into DAG maturity. It improves the state of knowledge on the sequencing of DAG by identifying six stages. Furthermore, the use of the critical path method is a novelty in the maturity literature. The use of maturity models to develop roadmaps is not new, but the idea of a critical path has hitherto failed to penetrate the research domain. The critical path method proved valuable in the identification of general stages. Therefore, the model can serve as a useful starting point for attaining a deeper understanding of maturity.

The study also contributes to practice. The results are valuable to those charged with instituting and maintaining DAG in organizations. The DAGM model enables them to determine the current maturity level of their organizations, to identify the level that is desired, and to plan progress. The DAGM model provides concrete means of configuring DAG in practice. It also highlights the importance of particular facets of data analytics maturity.

Future research should test the DAGM model through a longitudinal case study to determine whether the proposals that emerge from the model have the desired effect in practice. Furthermore, quantitative research on the effectiveness of the model can examine the relationship between maturity in DAG and data analytics performance.

Conclusion



This chapter summarizes the dissertation. Section 7.1 summarizes and discusses the main findings that pertain to the research sub-questions. Section 7.2 outlines the principal implications for theory and practice. Finally, Section 7.3 describes the limitations of the research and provides suggestions for future work.

7.1. Research questions and conclusions

The chief purpose of this dissertation is to contribute to the successful application of data analytics within organizations. It centered on the process perspective and on the governance perspective. Each perspective is relevant to the answers of multiple research questions, and the sections that follow discuss them in turn.

7.1.1. Process perspective

Research Question A.1: What are the existing process methodologies for guiding a data analytics project?

This question was addressed by a systematic literature review, which resulted in an overview of the state of the art. Suggested improvements to conventional KD process methodologies were also reviewed. The review focused on the contributions that have been published after the most recent review, which was conducted by Mariscal et al. (2010). The review presented here showed that the six main steps of the CRISP-DM model remain valid across all of the process methodologies that were identified. In addition, the review revealed that the application of conventional process methodologies to new contexts, such as a big data analytics or healthcare, is an essential driver of proposals for new process methodologies in the literature. The adjustments that have been advanced involve the introduction of steps and tasks to the CRISP-DM model (Angee, 2018; Grady, 2016; Li et al., 2016). Examples of these new steps include a problem formulation step and a maintenance step. However, the utility of these new steps is limited. An in-depth comparison with the CRISP-DM guide revealed that the new steps are not completely new but merely extensions or elaborations of old ideas. Another important finding of the literature review is that there is interest in the use of agile approaches in data analytics projects. These approaches improve data analytics processes by working iteratively. Different methods have been proposed to achieve this improvement. They include Scrum, pair programming, and continuous integration. However, in a data analytics project, it is difficult to implement an agile method that accords with the Scrum standards because the nature of data preparation obstructs planned sprints. This finding led to a follow-up research question.

Research Question A.2: How can the Scrum method be applied to improve the execution of data analytics projects in organizations?

This question was addressed by developing and evaluating a Scrum-based data analytics methodology. The methodology in question uses the CRISP-DM model as a baseline and adds Scrum roles, events, and artefacts. The evaluation of the

Scrum-based data analytics methodology in practice revealed that completing the data preparation step in a time-boxed sprint is problematic. The experts who were interviewed suggested affording more attention to the additional step in the redefined Scrum-based methodology. During the data preparation step, the development team often depends on the availability of data. This dependency can consume all the time that is available for a sprint. Consequently, it is difficult to execute the sprint event and to deliver incremental value. Therefore, the recommendation in the redefined Scrum-based data analytics methodology is to split the sprints into two separate types. Such split is commonly used in software development (Qureshi et al., 2012). In data analytics, one type of sprint focuses on data preparation and the other focuses on data modelling to deliver incremental value. This improvement to the Scrum-based data analytics methodology makes it more effective in projects with dedicated data preparation. However, the split is not useful in all project types, especially in data-centric projects. It is hard to plan such projects as sprints because the result is unclear. This finding calls for more research into the suitability of different process methodologies for particular project types.

Research Question A.3: How can different data analytics process methodologies support the execution of different types of data analytics projects?

This research question was addressed by a multiple case study. The results indicate that the frequency of deploying the project outcome is an important criterion in the selection of process methodologies. The case study yielded a framework that shows what process methodology is most useful when considering the envisioned deployment of the outcome. The results of this study are valuable for practitioners because they enable them to choose a process methodology that fits their data analytics projects. For example, practitioners could choose a more conventional process methodology when the deployment of the project outcome has a specific end and a shorter development cycle. In these circumstances, the team is assembled for a limited time, and the project ends when the allotted time expires or when the objective is fulfilled. A more iterative process model can be chosen when the deployment of the project outcome has a longer development cycle and no defined end. In such a situation, the ongoing flow of data needs to be analyzed, and the analysis would benefit from a more automated and maintainable process.

7.1.2. Governance perspective

Research Question B.1: What governance mechanisms can organizations use to govern their data analytics activities?

A multiple case study was conducted to answer this research question. A preliminary framework was developed. It drew on the extant data analytics governance literature, and it represents the first scientifically grounded framework for data analytics governance (DAG), as shown in Figure 7.1. This resulted in the DAG framework, which has two levels. The first level of the framework comprises three categories of governance mechanisms, namely structural, process, and relational. The second level contains more detailed sub-mechanisms in each of the three categories, resulting in a total of nine data analytics governance mechanisms. In the second step, the framework was evaluated within three organizations to elicit illustrations of the application of DAG mechanisms, providing a deeper understanding of the practical use of the sub-mechanisms. The empirical findings confirmed the existence of all nine governance sub-mechanisms. This contributes to the construct validity of the framework: at least one instantiation was found for each of the nine governance mechanisms that it proposes. An attempt was made to identify mechanisms that did not fit either of the nine mechanisms, but none emerged from the data. It appears, then, that the current set of mechanisms is sufficiently comprehensive. Therefore, it can serve as a guideline for DAG.

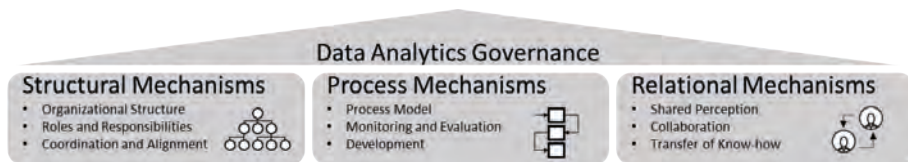


Figure 7.1. Data Analytics Governance Framework

Research Question B.2: How should organizations progress in the governance of their data analytics?

A maturity model for DAG was created to answer this question. It uses a design science research method. The model was designed, demonstrated, and evaluated in three cycles, resulting in the data analytics governance maturity (DAGM) model. Its evaluation in five organizations demonstrated its validity for measuring DAG. Furthermore, the model provided insights into the maturation of DAG at organizations by highlighting a critical path that passes through six stages: creating managerial awareness, establishing a structure for data analytics, measuring data analytics

activities, long-term planning, improving flexibility, and ensuring the continuity of data analytics practices. These stages provide concrete means to calibrate DAG and to prioritize tasks in the pursuit of maturity in DAG.

7.2. Implications

The sections that follow describe the implications of the research for theory and practice.

7.2.1. Theoretical implications

The ascent of data has intensified interest in its uses, including the use of process methodologies in data analytics projects. The last overview of process methodologies was published in 2010 (Mariscal et al., 2010). It showed that all process methodologies that were in circulation at that time overlapped to some extent and that it could be inferred that they were all instantiations of a general process. However, since 2010, the rise of data analytics has precipitated the emergence of new methods. Accordingly, Chapter 2 overviews the state of the art in data analytics process methodologies.

Chapter 2 revealed that data analytics process methodologies favor the use of agile methods. Using agile methods is also exceedingly popular in other areas, such as software engineering. However, transposing methods to new settings is a point of difficulty for many organizations. It is rare that a method can be used in a new setting without any adaptations. The Scrum-based data analytics method, as described in Chapter 3, is a clear example. Although the literature recommends it (do Nascimento & de Oliveira, 2012; Saltz, Shamshurin, & Crowston, 2017; Schmidt & Sun, 2018), it does not describe the use of Scrum clearly. Research on the adoption of Scrum in data analytics and the description of a Scrum-based data analytics method provides a better understanding of the artefact, the roles, and the events that are vital in such settings.

This said, the use of process methodologies remains a question of organizational context. Different organizations conduct different types of projects. The cases studies in Chapter 4 show that the process model can differ with project types. Although it was not possible to identify all project types, the need to recognize different process methodologies in different project types became apparent. No single process methodology fits all projects, and choosing the right model is a matter of organizational context (Umanath, 2003).

The growing interest in the use of data has also elevated the importance of governance. The existing models for the governance of IT and data are insufficiently comprehensive and fail to account for the value created by data analytics. The literature on DAG is limited to highlighting its necessity and the various problems that result from its absence. An examination of different DAG mechanisms was exigent. Thus, Chapter 5 introduced a novel typology of governance mechanisms in the data analytics literature and based a DAG framework on it. The framework extends the IT and data governance literature (De Haes & Van Grembergen, 2004; Tallon et al., 2013) by explaining how the nine generic governance mechanisms can be implemented in data analytics. The framework provides an overview of the elements of DAG, which makes it a solid foundation for further research in the field.

The framework in Chapter 5 was used further to develop the DAGM model in Chapter 6. Besides the practical use of that model in organizations, it also has scientific value: it explains how organizations mature in DAG, and it sheds light on the sequence in which DAG practices ought to be implemented. Therefore, the model can serve as useful starting point for further research on DAG maturity.

7.2.2. Implication for practice

For practitioners, this dissertation provides insights into the use of process methodologies in data analytics projects. It emerged that CRISP-DM is still used widely and without significant changes. It remains a helpful model for managers. Therefore, it was taken as a basis for developing an agile process methodology. This is done by using the agile method Scrum in CRISP-DM. To use that method, organizations should split their data analytics activities into a data preparation sprint and a data modelling sprint. Managers must also be sensitive to project type when selecting data analytics processes. For example, in a project where the outcome needs to be updated frequently, a more agile method is better. In this way, the research improves the use of process methodologies in data analytics projects.

The dissertation also shows how organizations can govern their data analytics. The DAG framework is valuable for organizations that wish to institute and maintain DAG. The different governance mechanisms of the framework provide concrete means of implementation. Moreover, the real-world cases provide concrete examples of the manifestations of these governance mechanisms. Organizations can use these examples to choose and customize a set DAG mechanisms from the framework, which they consider most appropriate for their organization. The dissertation also developed a maturity model. Organizations can use it to see where they are and where they want to go with their DAG. Consequently, the research contributes to mitigating problems in the governance of data analytics.

7.3. Limitations

The studies presented in this dissertation were executed as carefully as possible to ensure high-quality results. However, sacrifices and trade-offs are inevitable in any study. The choices that were made created limitations, which must be considered when the results of the research are used.

One such limitation stemmed from the employment of master's students as researchers. For example, the students conducted interviews to collect data for their theses. They are relatively inexperienced, and their approach to research varies. Several measures were taken to train and supervise them, including education about interview techniques (Saunders et al., 2009). Training took place in large-group sessions to ensure uniformity of approach and to provide a forum for discussing experiences and problems.

Another limitation is that, owing to its scope, the dissertation relies primarily on artificial evaluation rather than more naturalistic methods, due to the research scope. As a result, the research only validated the Scrum-based data analytics methodology and the DAGM model, and there was no room for implementing these artefacts in practice. This may cause some potential bias about the performance of the artefact. Implementation would have yielded additional insights about how these artefacts could be applied in practice. However, the validation of the artefacts made testing more efficient and yielded more immediate suggestions for improvement (Venable et al., 2016).

Finally, the interviewees were predominantly Dutch, as were the organizations where they worked. Dutch interviewees and organizations predominated because the research was conducted within the Center of Actionable Research of the Open University (CAROU), which is located on the Brightlands campus in Heerlen. Therefore, given the nature and scope of the study it is difficult to generalize the results international. Interviewing representatives of different nationalities could have led to different results. However, the lack of international differences are unlikely to have influenced the results on the governance framework, many of the organizations under observation are parts of large international companies. As a result, governance choices are often selected at international headquarters.

7.4. Future research

Two suggestions for future research on the process perspective can be made. The first suggestion focuses on the use of Scrum for a data analytics process methodology. A next step would be to do an extensive evaluation with Scrum experts after the validation in chapter 2 that was limited to data science experts. The use of Scrum experts allows to investigate if Scrum-DA can be used according to the Scrum principles. This could then be followed up with research on how Scrum-DA can be improved if it is applied to a real data analytics project and if its users are given an opportunity to reflect on their experiences. Such a study may take the form of action research where a real project team is observed while using Scrum-DA. The second suggestion in the process perspective would be to validate the framework of project types and process methodologies with more cases. It is important to test whether it is helpful for organizations in choosing the correct process methodology for the project that they plan to execute.

Future research on governance should circle around four considerations. First, it is necessary to validate the usefulness of the DAG framework theoretically. The nine different DAG mechanisms provide a broad perspective on the subject, but it is unclear if they can improve the governance of data analytics. For instance, future research could apply viable system model (VSM) theory to more in-depth case studies to describe and diagnose the approach of organizations to DAG and to examine the relation between the “essential elements of organization” in VSM. Second, the influence of contextual, or contingency, factors, such as the role of data analytics in an organization, should be investigated. Third, the maturity model should be tested in a practical longitudinal case study. Such a study can also ascertain whether the maturity of DAG develops alongside the DAGM model. Finally, quantitative research on the effectiveness of the data analytics maturity model is recommended. As this dissertation contributed to developing the nascent field of DAG, future quantitative research is needed (Edmondson & Mcmanus, 2007). Quantitative research will enable an examination of the relationship between maturity in DAG and the data analytics performance of organizations.

References

- Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016, Feb). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, 17(2), 1-XXXII.
- Ahangama, S., & Poo, D. C. C. (2014). Unified Structured Process for Health Analytics. *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, 8(11), 768-776.
- Ahangama, S., & Poo, D. C. C. (2015a). Designing a Process Model for Health Analytic Projects. PACIS 2015 Proceedings. 3.,
- Ahangama, S., & Poo, D. C. C. (2015b). What methodological attributes are essential for novice users to analytics? - an empirical study. International Conference on Human Interface and the Management of Information,
- Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment? *International Journal of Production Economics*, 182, 113-131.
- Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: an analysis of the literature. *Journal of Decision Systems*, 25(1), 64-75.
- Almeida, R., Pereira, R., & Mira da Silva, M. (2013). IT Governance Mechanisms: A Literature Review. International Conference on Exploring Services Science,
- Angee, S. (2018). Towards an Improved ASUM-DM Process Methodology for Cross-Disciplinary Multi-organization Big Data & Analytics Projects. International Conference on Knowledge Management in Organizations,
- Avery, A. A., & Cheek, K. (2015). Analytics Governance : Towards a Definition and Framework. Twenty-first Americas Conference on Information Systems,
- Ayele, W. (2020). Adapting CRISP-DM for Idea Mining. *International Journal of Advanced Computer Science and Applications*, 11(6), 20-32.
- Baijens, J., & Helms, R. (2019). Developments in Knowledge Discovery Processes and Methodologies: Anything New? Twenty-fifth Americas Conference on Information Systems,
- Baijens, J., Helms, R., & Iren, D. (2020). Applying Scrum in Data Science Projects. IEEE 22nd Conference on Business Informatics (CBI),
- Baijens, J., Helms, R., & Kusters, R. (2020). Data Analytics Project Methodologies: Which One to Choose? Proceedings of the 2020 International Conference on Big Data in Management,
- Baijens, J., Helms, R. W., & Velstra, T. (2020). Towards a Framework for Data Analytics Governance Mechanisms. Twenty-Eighth European Conference on Information Systems (ECIS2020),
- Barbour, J. B., Treem, J. W., & Kolar, B. (2018). Analytics and expert collaboration: How individuals navigate relationships when working with organizational data. *Human Relations*, 71(2), 256-284.
- Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing Maturity Models for IT Management. *Business & Information Systems Engineering*, 1(3), 213-222.
- Been, R., & Davenport, T. H. (2019). *Companies Are Failing in Their Efforts to Become Data-Driven*. <https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven>
- Berndtsson, M., Forsberg, D., Stein, D., & Svahn, T. (2018). Becoming a Data-Driven Organization. Twenty-Sixth European Conference on Information Systems,
- Boyd, A. (2011). *What is analytics?* <https://pubsonline.informs.org/doi/10.1287/LYTX.2011.02.09/full/>
- Brendel, A. B., Trang, S., Marrone, M., Lichtenberg, S., Brendel, A. B., & Marrone, M. (2020). What to do for a Literature Review ? – A Synthesis of Literature Review Practices. Americas Conference on Information Systems,
- Bruin de, T., Kulkarni, U., Freeze, R. D., & Rosemann, M. (2005). Understanding the Main Phases of Developing a Maturity Assessment Model. ACIS 2005 Proceedings,
- Burant, T. J., Gray, C., Ndaw, E., McKinney-Keys, V., & Allen, G. (2007). The Rhythms of a Teacher Research Group. *Multicultural Perspectives*, 9(1), 10-18.
- Chan, F. K. Y., & Thong, J. Y. L. (2009). Acceptance of agile methodologies : A critical review and conceptual framework. *Decision Support Systems*, 46(4), 803-814.

- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Crisp-Dm 1.0* (9780769532677). (CRISP-DM Consortium, Issue.
- Chen, H.-M., Schütz, R., Kazman, R., & Matthes, F. (2017). How Lufthansa Capitalized on Big Data for Business Model Renovation. *MIS Quarterly Executive*, *16*(1), 299-320.
- Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *Mis Quarterly*, *36*(4), 1165-1188.
- Darke, P., Shanks, G., & Broadbent, M. (1998). Successfully completing case study research: combining rigour, relevance and pragmatism. *Information Systems Journal*, *8*(4), 273-289.
- Das, M., Cui, R., Campbell, D. R., Agrawal, G., & Ramnath, R. (2015). Towards methods for systematic research on big data. Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015,
- Davenport, T. H. (2006). Competing on analytics. *Harvard business review*, *84*(1), 98-107.
- Davenport, T. H. (2013). Analytics 3.0. *Harvard Business Review*, *91*(12), 64-72.
- Davenport, T. H., Barth, P., & Bean, R. (2012). How ' Big Data ' is Different. *MIT Sloan Management Review*, *54*(1), 22-24.
- Davenport, T. H., Harris, J. G., Long, D. W., & Jacobson, A. L. (2001). Data to Knowledge to Results: Building an Analytic Capability. *California Management Review*, *43*(2), 117-138.
- Davenport, T. H., Mittal, N., & Saif, I. (2020). *What Separates Analytical Leaders From Laggards?* <https://sloanreview.mit.edu/article/what-separates-analytical-leaders-from-laggards/>
- De Haes, S., & Van Grembergen, W. (2004). IT Governance and its Mechanisms. *Information Systems Control Journal*, *1*, 27-33.
- De Haes, S., & Van Grembergen, W. (2009). An Exploratory Study into IT Governance Implementations and its Impact on Business/IT Alignment. *Information Systems Management*, *26*(2), 123-137.
- Delen, D., & Demirkan, H. (2013). Data , information and analytics as services. *Decision Support Systems*, *55*, 359-363.
- Delen, D., & Ram, S. (2018). Research challenges and opportunities in business analytics. *Journal of Business Analytics*, *1*(1), 2-12.
- Dharmapal, S. R., & Sikamani Thirunadana, K. (2016). Big data analytics using agile model. International Conference on Electrical, Electronics, and Optimization Techniques,
- do Nascimento, G. S., & de Oliveira, A. A. (2012). An Agile Knowledge Discovery in Databases Software Process. The Second International Conference on Advances in Information Mining and Management compliance,
- Drechsler, A., & Hevner, A. R. (2018). Utilizing , Producing , and Contributing Design Knowledge in DSR Projects. International Conference on Design Science Research in Information Systems and Technology,
- Dremel, C., Herterich, M. M., Wulf, J., & vom Brocke, J. (2018). Actualizing Big Data Analytics Affordances : A Revelatory Case Study. *Information & Management*, *57*(1).
- Dremel, C., Herterich, M. M., Wulf, J., Waizmann, J.-C., & Brenner, W. (2017). How Audi AG established big data analytics in its digital transformation. *MIS Quarterly Executive*, *16*(2), 81-100.
- Dul, J., & Hak, T. (2008). *Case Study Methodology in Business Research*. <https://doi.org/10.1007/s13398-014-0173-7>
- Edmondson, A. M. Y. C., & Mcmanus, S. E. (2007). Methodological Fit in Management Field Research. *Academy of Management Review*, *32*(4), 1155-1179.
- Espinosa, J. A., & Armour, F. (2016). The big data analytics gold rush: A research framework for coordination and governance. Proceedings of the Annual Hawaii International Conference on System Sciences.
- EYGM-Limited. (2015). *Becoming an analytics-driven organization to create value A report in collaboration with Nimbus Ninety*. [https://www.ey.com/Publication/vwLUAssets/EY-global-becoming-an-analytics-driven-organization/\\$FILE/ey-global-becoming-an-analytics-driven-organization.pdf](https://www.ey.com/Publication/vwLUAssets/EY-global-becoming-an-analytics-driven-organization/$FILE/ey-global-becoming-an-analytics-driven-organization.pdf)
- Fadler, M., & Legner, C. (2021). Toward big data and analytics governance : redefining structural governance mechanisms. Proceedings of the 54th Hawaii International Conference on System Sciences.

- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). Knowledge Discovery and Data Mining: Towards a Unifying Framework. Int Conf on Knowledge Discovery and Data Mining,
- Félix, B. M., Tavares, E., & Cavalcante, N. W. F. (2018). Critical success factors for Big Data adoption in the virtual retail: Magazine Luiza case study. *Review of Business Management*, 20(1), 112-126.
- Gao, J., Koronios, A., & Selle, S. (2015). Towards A Process View on Critical Success Factors in Big Data Analytics Projects. Twenty-first Americas Conference on Information Systems,
- George, G., Haas, M. R., & Pentland, A. (2014). Big Data and Management. *Academy of Management Journal*, 57(2), 321-326.
- Ghasemaghahi, M., Ebrahimi, S., & Hassanein, K. (2018). Data analytics competency for improving firm decision making performance. *Journal of Strategic Information Systems*, 27(1), 101-113.
- Grady, N. W. (2016). Knowledge Discovery in Data Science KDD meets Big Data. IEEE International Conference on Big Data,
- Grady, N. W., Payne, J. A., & Parker, H. (2017). Agile big data analytics: AnalyticsOps for data science. Proceedings 2017 IEEE International Conference on Big Data, stein.
- Gröger, C. (2018). Building an Industry 4.0 Analytics Platform. *Datenbank-Spektrum*, 18(5), 5-14.
- Grossman, R. L. (2018). A framework for evaluating the analytic maturity of an organization. *International Journal of Information Management*, 38(1), 45-51.
- Grossman, R. L., & Siegel, K. P. (2014). Organizational Models for Big Data and Analytics. *Journal of Organization Design*, 3(1), 20-25.
- Grover, V., Chiang, R. H. L., Liang, T.-p., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics : A Research Framework. *Journal of Management Information Systems*, 35(2), 388-423.
- Gudivada, V. N. (2017). Data analytics: fundamentals. In M. Chowdhury, A. Apon, & K. Dey (Eds.), *Data Analytics for Intelligent Transportation Systems* (pp. 31-67). NY: Elsevier.
- Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *Journal of Strategic Information Systems*, 26(3), 191-209.
- Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information and Management*, 53(8), 1049-1064.
- He, J., & Mahoney, J. (2006). Firm capability, corporate governance, and firm competitive behavior: A multi-theoretic framework. *International Journal of Strategic Change Management*, 1(4), 293-318.
- Hevner, A. R., March, S. T., Park, J., Ram, S., & Ram, S. (2004). Design Science in Information Systems Research. *Mis Quarterly*, 28(1), 75-105.
- Hofstede, G. (1981). Culture and Organizations. *International Studies of Management & Organization*, 10(4), 15-41.
- Howard, C. E. (2016). *Lufthansa Technik debuts predictive maintenance, condition monitoring platform to avoid failures, optimize MRO*. <https://www.intelligent-aerospace.com/commercial/article/16538912/lufthansa-technik-debuts-predictive-maintenance-condition-monitoring-platform-to-avoid-failures-optimize-mro>
- Jackman, M., & Reddy, M. (2020). *Analytics at Netflix: Who We Are and What We Do*. <https://netflixtechblog.com/analytics-at-netflix-who-we-are-and-what-we-do-7d9c08fe6965>
- Jakobsen, C. R., & Johnson, K. A. (2008). Mature agile with a twist of CMMI. Agile 2008 Conference,
- Jansen, S. (2020). A focus area maturity model for software ecosystem governance. *Information and Software Technology*, 118.
- Jensen, M. H., Nielsen, P. A., & Persson, J. S. (2019). Managing Big Data Analytics Projects: The Challenges of Realizing Value. Proceedings of the 27th European Conference on Information Systems (ECIS),
- Jeske, M., Gruner, M., & Weiß, F. (2013). *Big data in logistics*. http://www.nsuchaud.fr/wp-content/uploads/2016/10/CSI_Studie_BIG_DATA_FINAL-ONLINE.pdf
- Kasten, J. E. (2020). Trust , Organizational Decision- Making , and Data Analytics : An Exploratory Study. *international journal of business intelligence research*, 11(1), 22-37.

- Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 53(1), 148-152.
- Kiron, D., Prentice, P., & Boucher Ferguson, R. (2014). The Analytics Mandate. *MIT Sloan Management Review*, 55(4), 1.
- Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., & Haydock, M. (2011). Analytics: The Widening Divide advantage through analytics. *MIT Sloan Management Review*, 53(2), 1-21.
- Kloör, B., Monhof, M., Beverungen, D., & Braäer, S. (2018). Design and evaluation of a model-driven decision support system for repurposing electric vehicle batteries. *European Journal of Information Systems*, 27(2), 171-188.
- Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies Mete Group,
- Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36(5), 700-710.
- Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 52(2), 21-32.
- Li, Y., Thomas, M. A., & Osei-Bryson, K.-M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems*, 91, 1-12.
- Luo, J., Wu, Z., Huang, Z., & Wang, L. (2016). Relational IT governance, its antecedents and outcomes: A study on Chinese firms. 2016 International Conference on Information Systems (ICIS),
- Mahoney, J. (2018, Jun). The Alliance for Innovation in Maternal Health Care: A Way Forward. *Clinical Obstetrics and Gynecology*, 61(2), 400-410.
- Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2009). Toward data mining engineering: A software engineering approach. *Information Systems*, 34(1), 87-107.
- Marbán, Ó. a., Mariscal, G. b., Menasalvas, E. a., & Segovia, J. a. (2007). An engineering approach to data mining projects. International Conference on Intelligent Data Engineering and Automated Learning, Berlin.
- Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*, 25(2), 137-166.
- Marr, B. (2017). *The Amazing Ways Spotify Uses Big Data, AI And Machine Learning To Drive Business Success*. <https://www.forbes.com/sites/bernardmarr/2017/10/30/the-amazing-ways-spotify-uses-big-data-ai-and-machine-learning-to-drive-business-success/?sh=49988a34bd2f>
- Martínez-Plumed, F., Contreras-ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. A. (2019). CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*,
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big Data : The Management Revolution. *Harvard Business Review*, 90(3), 60-68.
- McNaughton, M., Rao, L., & Mansingh, G. (2017). An agile approach for academic analytics: a case study. *Journal of Enterprise Information Management*, 30(5), 701-722.
- McShea, C., Oakley, D., & Mazzei, C. (2016). *The Reason So Many Analytics Efforts Fall Short*. <https://hbr.org/2016/08/the-reason-so-many-analytics-efforts-fall-short>
- Mikalef, P., Krogstie, J., Mikalef, P., Danielsen, F., & Olsen, D. H. (2017). Big Data Analytics Capability : Antecedents and Business Value. Twenty First Pacific Asia Conference on Information Systems,
- Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2017). Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and E-Business Management*, 16(3), 547-578.
- Mortenson, M. j., Doherty, N. f., & Robinson, S. (2015). Operational research from Taylorism to Terabytes_ A research agenda for the analytics age. *European Journal of Operational Research*, 241(3), 583-595.
- Muntean, M., & Surcel, T. (2013). Agile BI – The Future of BI. *Informatica Economică*, 17(3), 114-124.
- Najafi, M., & Engineer, U. E. (2008). Two Case Studies of User Experience Design and Agile Development. Agile 2008 Conference,
- Niño, H. A. C., Niño, P. C. J., & Ortega, R. M. (2020). Business intelligence governance framework in a university : Universidad de la costa case study. *International Journal of Information Management*, 50, 405-412.

- Oestreich, T. (2016). *Establish a Framework for Analytics Governance* (Gartner Business Intelligence and Analytics Summit, Issue).
- Okoli, C., & Schabram, K. (2012). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *SSRN Electronic Journal*, 10.
- Otto, B. (2011). A Morphology of the Organisation of Data Governance. ECIS 2011 Proceedings,
- Pappas, I. O., Mikalef, P., Giannakos, M. N., Krogstie, J., & 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, 479-491.
- Parkins, D. (2017). *The world's most valuable resource is no longer oil, but data*. Economist, The. <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>
- Pearlson, K. E., & Saunders, C. S. (2013). *Managing & Using Information Systems: A Strategic Approach*.
- Peppers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45-77.
- Pöppelbuß, J., & Röglinger, M. (2011). What makes a useful maturity model? A framework of general design principles for maturity models and its demonstration in business process management. ECIS 2011 Proceedings,
- Power, D. J., Heavin, C., McDermott, J., & Daly, M. (2018). Defining business analytics: an empirical approach. *Journal of Business Analytics*, 1(1), 40-53.
- Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact evaluation in information systems design science research - A holistic view. Proceedings - Pacific Asia Conference on Information Systems, PACIS 2014,
- Provost, F., & Fawcett, T. (2013a). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1), 51-59.
- Provost, F., & Fawcett, T. (2013b). *Data Science for Business*. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Qureshi, M. R. J., Barnawi, A., & Ahmad, A. (2012). Proposal of Implicit Coordination Model for Performance Enhancement Using Sprint Zero. *I.J. Information Technology and Computer Science*, 4(9), 45-52.
- Rau, K. G. (2004). Effective governance of it: Design objectives, roles, and relationships. *Information Systems Management*, 21(4), 35-42.
- Röglinger, M., Pöppelbuß, J., & Becker, J. (2012). Maturity Models in Business Process Management. *Business Process Management Journal*, 18.
- Rose, R. (2016). *Defining analytics: a conceptual framework*. <https://www.informs.org/ORMS-Today/Public-Articles/June-Volume-43-Number-3/Defining-analytics-a-conceptual-framework>
- Saldaña, J. (2015). *The coding manual for qualitative researchers*. Sage. <https://doi.org/10.1017/CBO9781107415324.004>
- Saltz, J. (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,
- Saltz, J. (2018). Identifying the key drivers for teams to use a data science process methodology. 26th European Conference on Information Systems,
- Saltz, J., Heckman, R., & Shamshurin, I. (2017). Exploring How Different Project Management Methodologies Impact Data Science Students. Twenty-Fifth European Conference on Information Systems (ECIS), Guimarães, Portugal,
- Saltz, J., & 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,
- Saltz, J., Shamshurin, I., & Connors, C. (2017). Predicting data science sociotechnical execution challenges by categorizing data science projects. *Journal of the Association for Information Science and Technology*, 68(12), 2720-2728.

- Saltz, J., Shamshurin, I., & Crowston, K. (2017). Comparing Data Science Project Management Methodologies via a Controlled Experiment. Proceedings of the 50th Hawaii International Conference on System Sciences,
- Saltz, J., & Sutherland, A. (2019). SKI : An Agile Framework for Data Science. 2019 IEEE International Conference on Big Data (Big Data),
- Saltz, J., & Sutherland, A. (2020). SKI : A New Agile Framework that supports DevOps , Continuous Delivery , and Lean Hypothesis Testing. Proceedings of the 53rd Hawaii International Conference on System Sciences.,
- Saltz, J., Wild, D., Hotz, N., & Stirling, K. (2018). Exploring Project Management Methodologies Used Within Data Science Teams. Twenty-fourth Americas Conference on Information Systems, New Orleans, 2018,
- Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research Methods for Business Students*. Pearson Education LTD. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Schmidt, C., & Sun, W. N. (2018). Synthesizing Agile and Knowledge Discovery: Case Study Results. *Journal of Computer Information Systems*, 58(2), 142-150.
- Schüritz, R., Brand, E., Satzger, G., & Bischoffshausen, J. (2017). How To Cultivate Analytics Capabilities Within an Organization ? – Design and Types of Analytics Competency Centers. Proceedings of the 25th European Conference on Information Systems (ECIS),
- Schwaber, K., & Sutherland, J. (2017). *The Scrum Guide: The Definitive The Rules of the Game* (9788578110796). (Scrum.Org, Issue. <https://scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf>
- Seddon, P. B., Constantinidis, D., Tamm, T., & Dod, H. (2017, May). How does business analytics contribute to business value? *Information Systems Journal*, 27(3), 237-269.
- Sharma, S. (2012). *An Integrated Knowledge Discovery and Data Mining Process Model* Virginia Commonwealth University].
- Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263-286.
- Smits, D. (2015). IT Governance Maturity : Developing a Maturity Model using the Delphi Method. 48th Hawaii International Conference on System Sciences,
- Smits, D., & Hillegersber van, J. (2017). The development of a hard and soft IT governance assessment instrument. *Procedia Computer Science*,
- Spruit, M., & Röling, M. (2014). ISFAM: the information security focus area maturity model. Twenty Second European Conference on Information Systems,
- Strauss, A. L. (1987). *Qualitative Analysis for Social Scientists*. Cambridge university press.
- Tallon, P. P. (2013). Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost. *IEEE computer society*, 46(6), 32-38.
- Tallon, P. P., Ramirez, R. V., & Short, J. E. (2013). The Information Artifact in IT Governance: Toward a Theory of Information Governance. *Journal of Management Information Systems*, 30(3), 141-178.
- Team, C. P. (2006). *Capability maturity model@ integration, version 1.2*.
- Umanath, N. S. (2003). The concept of contingency beyond “It depends”: illustrations from IS research stream. *Information & Management*, 40(6), 551-562.
- Van Steenberg, M., Bos, R., Brinkkemper, S., & Weerd, I. V. D. (2010). The Design of Focus Area Maturity Models. International Conference on Design Science Research in Information Systems,
- Van Steenberg, M., Van den Berg, M., & Brinkkemper, S. (2007). A Balanced Approach to Developing the Enterprise Architecture Practice. International Conference on Enterprise Information Systems,
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, 25(1), 77-89.
- Viaene, S., & Bunder, A. V. d. (2011). The secrets to managing business analytics projects. *MIT Sloan Management Review*, 53(1), 65-69.
- Vidgen, R., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research*, 261(2), 626-639.

- Volk, M., Jamous, N., & Turowski, K. (2017). Requirements Engineering for Big Data Projects Ask the Right Questions: Requirements Engineering for the Execution of Big Data Projects. Twenty-third Americas Conference on Information Systems,
- vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In *Design Science Research. Cases*. <https://doi.org/10.1007/978-3-030-46781-4>
- Vosloo, P., & Naidoo, R. (2019). Contextual critical success factors for the implementation of business intelligence & analytics : A qualitative case study. CONF-IRM 2019 Proceedings,
- Walker, J. (2017). *Big data strategies disappoint with 85 percent failure rate*. <http://www.digitaljournal.com/tech-and-science/technology/big-data-strategies-disappoint-with-85-percent-failure-rate/article/508325>
- Ward, J. S., & Barker, A. (2013). *Undefined By Data: A Survey of Big Data Definitions* arXiv preprint arXiv:1309.5821,
- Watson, H. J. (2014). Tutorial : Big Data Analytics : Concepts , Technologies , and Applications. *Communications of the ACM*, 34(65), 1247-1268.
- Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer*, 40(9), 96-99.
- Weber, K., Otto, B., & Osterle, H. (2009). One Size Does Not Fit All — A Contingency Approach to Data Governance. *ACM Journal of Data and Information Quality*, 1(1), 1-27.
- Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future : Writing a Literature Review. *Mis Quarterly*, 26(2), 13-23.
- Wegener, R., & Sinha, V. (2013). *The value of Big Data: How analytics differentiates winners*. <https://www.bain.com/insights/the-value-of-big-data>
- Weill, P., & Ross, J. (2005). A Matrixed Approach to Designing IT Governance. *MIT Sloan Management Review*, 46(2), 26-34.
- Weill, P., & Ross, J. W. (2004). *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results*. Harvard Business School Press.
- White, A. (2019). *Our Top Data and Analytics Predicts for 2019*. https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/
- Wieringa, R. J. (2014). *Design Science Methodology for information systems and software engineering*. Springer Berlin Heidelberg.
- Williams, L. (2010). Agile Software Development Methodologies and Practices. *Advances in Computers*, 80, 1-44.
- Willis, R. J. (1985). Critical path analysis and resource constrained project scheduling - Theory and practice. pdf. *Journal of Operational Research*, 21(2), 149-155.
- Wixom, B. H., Yen, B., & Relich, M. (2013). Maximizing Value from Business Analytics. *MISQ Executive*, 12(2), 111-123.
- Wu, P.-J., Straub, D. W., & Liang, T.-P. (2015). How information technology governance mechanisms and strategic alignment influence organizational performance: Insights from a matched survey of business and IT managers. *Mis Quarterly*, 39(2), 497-518.
- Yamada, A., & Peran, M. (2018). Governance framework for enterprise analytics and data. 2017 IEEE International Conference on Big Data,
- Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a Network Highway for Big Data: Architecture and Challenges. *Ieee Network*, 28(4), 5-13.
- Yin, R. k. (2017). *Case study research and applications: Design and methods*. SAGE Publications.
- Zogaj, S., & Bretschneider, U. (2014). Analyzing governance mechanisms for crowdsourcing information systems: A multiple case analysis. Proceedings European Conference on Information Systems,

Summary

This dissertation contributes to a solution for organizational problems in data analytics. For this purpose, the dissertation is divided into two parts. First, it investigates contributions to solutions from a process perspective. Second, it investigates contributions to solutions from a governance perspective.

Chapters 2, 3 and 4 focus on a process perspective. To understand the use of processes in data analytics, chapter 2 created an overview of data analytics process methodologies by reviewing the research into data analytics process methodologies since 2010. At that moment the ascent of data has intensified interest in organizations. The review revealed that the application of conventional process methodologies to new contexts, such as big data analytics or healthcare, is an essential driver of proposals for new process methodologies in the literature. Another important finding of the literature review is that there is interest in the use of agile approaches, such as the Scrum method.

Although the literature recommends using Scrum for data analytics, it does not describe the use of Scrum clearly. Therefore, the research on the adoption of Scrum in data analytics and the description of a Scrum-based data analytics method is necessary. The research in chapter 3 integrated the Scrum method in the CRISP-DM methodology using a Design Science Research approach and provided a better understanding of the artefact, the roles, and the events that are vital in such settings. This new methodology was then evaluated using expert interviews. Analysis of the expert interviews resulted in a further refinement of the Scrum data analytics methodology.

However, the use of Scrum for a data analytics process methodology was not a solution for all types of projects. Therefore, the research in chapter 4 provided a structured description that helps to determine what type of process methodology works for different types of data analytics projects. More specifically, by grouping different project characteristics it was possible to identify the most appropriate process methodology for a specific type of project. The results of six different case studies show that continuous projects would benefit from an iterative methodology. Although it was not possible to identify all project types, the need to recognize different process methodologies in different project types became apparent. No single process methodology fits all projects, and choosing the right model is a matter of organizational context.

The second perspective is addressed in chapters 5 and 6, and focuses on the governance perspective. The growing interest in the use of data has also elevated the importance of governance. The existing models for the governance of IT and data are insufficiently comprehensive and fail to account for the value created by data analytics. The literature on data analytics governance is limited to highlighting its necessity and the various problems that result from its absence. An examination of different data analytics governance mechanisms was exigent. In chapter 5 data analytics governance mechanisms were identified to better understand how data analytics governance can be achieved. To this end, a literature review was conducted to identify a preliminary framework. The framework was validated, and extended, in three case studies by identifying practical implementations of governance mechanisms. This resulted in a novel typology data analytics governance mechanisms describing several structural, process and relational mechanisms. This framework can assist managers in designing data analytics governance mechanisms for their specific organization and provides a solid foundation for further research in the field.

Based on the typology in chapter 5, chapter 6 builds an artefact to assess the maturity for data analytics governance. The development of the maturity model provided a deeper understanding of the governance of data analytics and its trajectory to maturity. The results reflect three cycles of design, demonstration, and evaluation. The final cycle involved summative evaluation and showed that the model was perceived as a valid measurement of data analytics governance maturity. This evaluation added to the knowledge of data analytics governance and culminated in a roadmap for the governance of data analytics. Furthermore, a critical path was identified. It consists of creating awareness, structuring, measuring, long-term planning, adapting, and continuing. The maturity model helped to explain how organizations mature in data analytics governance, and it sheds light on the sequence in which data analytics governance practices ought to be implemented. Therefore, the model can serve as a useful starting point for further research on data analytics governance maturity.

Samenvatting (Dutch)

Dit proefschrift draagt bij aan een oplossing voor organisatorische problemen in data analytics. Daarom is het proefschrift opgesplitst in twee delen. Ten eerste onderzoekt het bijdragen aan oplossingen vanuit een procesperspectief. Ten tweede onderzoekt het vanuit een governance perspectief.

De hoofdstukken 2, 3 en 4 richten zich op het procesperspectief. Om het gebruik van processen in data analytics te begrijpen, is in hoofdstuk 2 een overzicht gemaakt van data analytics procesmethodologieën door onderzoeken sinds 2010 te reviewen. Vanaf dat moment heeft de opkomst van data de belangstelling in organisaties geïntensiveerd. Uit de review bleek dat de toepassing van conventionele procesmethodologieën op nieuwe contexten, zoals big data analytics of gezondheidszorg, een essentiële aanjager is van voorstellen voor nieuwe procesmethodologieën in de literatuur. Een andere belangrijke bevinding van het literatuuronderzoek is dat er belangstelling is voor het gebruik van agile benaderingen, zoals de Scrum-methode.

Hoewel in de literatuur het gebruik van Scrum voor data analytics wordt aanbevolen, wordt het gebruik van Scrum niet duidelijk beschreven. Daarom is onderzoek naar de adoptie van Scrum in data analytics en de beschrijving van een op Scrum gebaseerde data analytics methode noodzakelijk. Het onderzoek in hoofdstuk 3 integreerde de Scrum methode in de CRISP-DM methodologie met behulp van een Design Science Research benadering en zorgde voor een beter begrip van het artefact, de rollen en de gebeurtenissen die van vitaal belang zijn in dergelijke setting. Deze nieuwe methodologie werd vervolgens geëvalueerd aan de hand van expertinterviews. Dit resulteerde in een verfijning van de Scrum data analytics methodologie.

Het gebruik van Scrum voor een data analytics proces methodologie was echter niet een oplossing voor alle soorten projecten. Daarom heeft het onderzoek in hoofdstuk 4 een gestructureerde beschrijving opgeleverd die helpt om te bepalen welk type procesmethodologie werkt voor verschillende typen data analytics projecten. Meer specifiek, door het groeperen van verschillende projectkenmerken was het mogelijk om de meest geschikte procesmethodologie voor een specifiek type project te identificeren. De resultaten van zes verschillende casestudies toonden aan dat continue projecten baat zouden hebben bij een iteratieve methodologie. Hoewel het niet mogelijk was alle projecttypes te identificeren, werd het duidelijk dat verschillende procesmethodologieën in verschillende projecttypes moeten worden erkend. Geen enkele procesmethodologie past bij alle projecten en het kiezen van het juiste model is een kwestie van organisatorische context.

De groeiende belangstelling voor het gebruik van gegevens heeft ook het belang van governance doen toenemen. De bestaande modellen voor de governance van IT en data zijn onvoldoende omvattend en houden geen rekening met de waarde die door data analytics wordt gecreëerd. De literatuur over data analytics governance beperkt zich tot het benadrukken van de noodzaak en de verschillende problemen die het gevolg zijn van het ontbreken ervan. Een onderzoek naar verschillende mechanismen voor de governance van data analytics was dan ook dringend gewenst. Daarom richten hoofdstuk 5 en 6 zich op het governance perspectief. In hoofdstuk 5 werden mechanismen voor data analytics governance geïdentificeerd om beter te begrijpen hoe data analytics governance kan worden bereikt. Daartoe werd een literatuurstudie uitgevoerd om een voorlopig raamwerk te identificeren. Het raamwerk werd gevalideerd en uitgebreid in drie casestudies door praktische implementaties van governance-mechanismen te identificeren. Dit resulteerde in een nieuwe typologie van mechanismen voor data analytics governance, die verschillende structurele, proces- en relationele mechanismen beschrijft. Dit raamwerk kan managers helpen bij het ontwerpen van mechanismen voor data analytics governance voor hun specifieke organisatie en biedt een solide basis voor verder onderzoek op dit gebied.

Op basis van de typologie in hoofdstuk 5, werd in hoofdstuk 6 een artefact ontwikkeld om de volwassenheid van data analytics governance te beoordelen. De ontwikkeling van het maturiteitsmodel heeft geleid tot een dieper inzicht in de governance van data analytics en het traject naar volwassenheid. De resultaten weerspiegelden drie cycli van ontwerp, demonstratie en evaluatie. De laatste cyclus omvatte een summatieve evaluatie en toonde aan dat het model werd gezien als een geldige meting van de volwassenheid van data analytics governance. Deze evaluatie droeg bij aan de kennis over data analytics governance en leidde tot een stappenplan voor de governance van data analytics. Bovendien werd een kritisch pad geïdentificeerd. Het bestaat uit bewustwording, structureren, meten, langetermijnplanning, aanpassen en doorgaan. Het volwassenheidsmodel heeft geholpen om te verklaren hoe organisaties verbeteren in de governance van data analytics. Ook geeft het inzicht op de volgorde waarin praktijken voor de governance van data analytics moeten worden geïmplementeerd. Het model kan daarom dienen als een nuttig startpunt voor verder onderzoek naar de volwassenheid van data analytics governance.

Acknowledgements

This dissertation would not have been accomplished without the assistance of several individuals.

First of all, I would like to thank my supervisor and promoter Remko Helms for all the work and attention he put into making my PhD project a success. He always created a nice working atmosphere that motivated me to finish my dissertation. His guidance provided a strong foundation for me to conduct my research. I also appreciate the patience and time he put into reviewing my writing pieces. Furthermore, I enjoyed the intense discussions we had about research, but also the discussions about the shared passion for the sport of basketball.

I would also like to thank my other supervisor Rob Kusters for his efforts in completing my PhD. He was an inspiration for continuing the research and kept track of whether the general outlines of the research were going in the right direction. Also, his knowledge and experience contributed to the success of my dissertation.

Other people I would like to thank are my co-authors Tjeerd, Deniz and Tim from my various publications for the contributions they made in my research.

Furthermore, I would like to thank my colleagues from CAROU for their input and the nice working atmosphere: Gerard, Martine, Alex, Stefano, Elianne, Mark, Petru, Wiebke and Lyana. Also, I would like to thank my colleagues from the Information Science department and the colleagues I got to know during the BISS period. Next to colleges, I would also like to thank the three student groups that contributed to the dissertation.

Besides the help I received from my work sphere, I would also like to thank the people in my private sphere. My parents, family and close friends from whom I always received a lot of support, but who also showed interest in the progress of my PhD.

I would especially like to thank my mother who not only contributed to the design of the dissertation but also continued to support me throughout my academic career. From the time I was in elementary school until now, she never pressured, but rather motivated me to achieve my goals. Something that I have always experienced as very pleasant.

Finally, of course, I want to thank my girlfriend Simone. With whom I have shared this whole process from beginning to end. She has always supported me when I needed it. She has always been understanding at times when I had to put a little more time into my research. Even when it was my fault for waiting until the last moment of the deadline. From now on, I have no more excuses not to pick up the vacuum cleaner for once.

Funding sources

The research in this dissertation was supported by the Province of Limburg, The Netherlands, under grant number SAS-2020-03117

About the author

Jeroen Baijens is born on the 20th of January 1992 in Stein. After completing his vmbo-tl and havo exams at secondary school, he started in 2011 with a bachelor industrial engineering at Zuyd University of applied sciences in Heerlen. He received his bachelor's degree in 2015 and started his master's degree in International business with a focus on supply chain management at the University of Maastricht, which he completed in 2017. During his studies he worked as a student assistant at Sitech Services, drawing up analysis reports.

From 2017, he started full time with his PhD research under the supervision of Prof. Dr. ir. Remko Helms at the department of Information Science at the faculty of science of The Open Universiteit the Netherlands. His research interests during this period included data analytics process methodologies and data analytics governance. Part of his research is conducted for the Center of Actionable Research of the Open University (CAROU), which is specialized in data science, artificial intelligence and social innovation. His work appeared in different Information Systems conferences.

Currently Jeroen is working as an assistant professor of the Open University for CAROU. In this position he continues the research he started in his PhD.

Email address: jeroen.baijens@ou.nl

Website: www.ou.nl/carou

List of publications

Baijens, J., & Helms, R. (2019). Developments in Knowledge Discovery Processes and Methodologies: Anything New? Twenty-fifth Americas Conference on Information Systems.

Moonen, N., Baijens, J., Ebrahim, M., & Helms, R. (2019) Small Business, Big Data: An assessment framework for (big) data analytics capabilities in SMEs. Academy of Management Proceedings.

Baijens, J., Helms, R., & Iren, D. (2020). Applying Scrum in Data Science Projects. IEEE 22nd Conference on Business Informatics (CBI).

Baijens, J., Helms, R., & Kusters, R. (2020). Data Analytics Project Methodologies: Which One to Choose? Proceedings of the 2020 International Conference on Big Data in Management.

Baijens, J., Helms, R. W., & Velstra, T. (2020). Towards a Framework for Data Analytics Governance Mechanisms. Twenty-Eighth European Conference on Information Systems (ECIS2020).

Baijens, J., Huygh, T., & Helms, R. W. (2021): Establishing and theorising data analytics governance: a descriptive framework and a VSM-based view, *Journal of Business Analytics*, DOI: 10.1080/2573234X.2021.1955021



Open Universiteit

CAROU