MASTER'S THESIS

Hoe het gebruik van de juiste processtappen leidt tot succesvollere data science projecten

van Krieken, D.

Award date:
2021

Hoe het gebruik van de juiste processtappen leidt tot succesvollere data science projecten

How the use of the correct process steps leads to more successful data science projects

Opleiding:	Open Universiteit, faculteit Management, Science & Technology Masteropleiding Business Process Management & IT
Programme:	Open University of the Netherlands, faculty of Management, Science & Technology Master Business Process Management & IT
Cursus:	IM0602 Voorbereiden Afstuderen BPMIT IM9806 Afstudeeropdracht Business Process Management and IT
Student:	Dirk van Krieken
Identiteitsnummer:	
Datum:	2021-09-14
Afstudeerbegeleider	Jeroen Baijens
Meelezer	Remko W. Helms
Derde beoordelaar	nvt
Versie nummer:	20210914
Status:	definitief

Abstract

In de eenentwintigste eeuw wordt door steeds technologieën een enorme hoeveelheid data gegenereerd. Data science helpt organisaties uit al deze data kennis en informatie te halen Om de beste waarde uit data science projecten te krijgen zijn er afgelopen jaren diverse projectmethoden ontwikkeld om deze projecten in te richten, waaronder de bekende CRISP-DM methode. Helaas is gebleken dat een groot deel van deze projecten niet succesvol is. Daarnaast zien we ook dat bij veel projecten, projectmethoden niet of gedeeltelijk worden toegepast. In dit onderzoek is onderzocht welke projectmethoden en processtappen gebruikt kunnen worden voor het succesvol afronden van data science projecten, dit heeft geresulteerd in een data science project framework. Dit framework laat zien welke projectkarakteristieken invloed hebben op de keuze van de juiste processtappen. Het framework is in een casestudy doormiddel van expertsinterviews en focusgroep interviews getoetst bij een middelgrote onderneming in de tech-sector. Als bijdrage aan de wetenschap wordt een overzicht gegeven welke factoren belangrijk zijn bij de keuze en inrichting van de processtappen en de projectmethode van een data science project.

Sleutelbegrippen

Data science, data analyse, projectmethode, processtappen, projectkarakteristieken

Samenvatting

Uit onderzoek is gebleken dat een groot gedeelte van de data science projecten niet succesvol was. Dit is bijzonder spijtig aangezien data science organisaties helpt om de juiste kennis en informatie uit de enorme hoeveelheid data welke tegenwoordig beschikbaar is te halen. Uit vervolgonderzoek is verder gebleken dat maar liefst 82% van de ondervraagde data sciencisten aangaf geen of slechts gedeeltelijk gebruik te maken van de beschikbare projectmethoden voor data science projecten, dit terwijl 85% aangaf te verwachten dat het uitvoeren van een betere procesaanpak tot betere projectresultaten zal leiden. Deze thesis is een onderzoek naar de invloed van de verschillende eigenschappen, de projectkarakteristieken van een data science project op de keuze voor de beste processtappen en/of projectmethoden. Voor dit onderzoek is gebruik gemaakt van een Design Science Research (DSR). Het doel is een framework te creëren welke inzicht en overzicht geeft van de projectkarakteristieken welke invloed hebben op de processtappen en daarmee de projectaanpak van een data science project. Dit framework kan hiermee bijdragen aan een succesvolle(re) aanpak van data science projecten. Het onderzoek geeft antwoord op de volgende onderzoeksvraag:

'Hoe kunnen projectmethoden/processtappen worden gebruikt om een data science project succesvol te maken?'

In het literatuuronderzoek is allereerst een overzicht weergegeven van de projectmethoden en de processtappen voor data science projecten. Als basis is het CRISP-DM model genomen, deze wordt in de praktijk nog steeds toegepast. Deze is aangevuld met een aantal processtappen, namelijk Life cycle selection Problem formulating (als aanvullend onderdeel op de stap business understanding), Conceptualization, Automate en Support / Maintenance, plus procesthema's, namelijk iteratiemogelijkheden en Agile methodieken als Scrum en Kanban, om de ontwikkelingen van data science project methoden weer te geven. Vervolgens is onderzocht welke projectkarakteristieken voor data science projecten worden onderscheiden. Deze zijn ingedeeld in vijf categorieën, namelijk projecttypen, Data context, Analytical context, Team context en Organizational context.
Er is niet één one-fits-all methode voor alle data science projecten, de keuze voor de juiste projectmethode en processtappen hangt af van de eigenschappen van het data science project. Deze afhankelijkheid is weergegeven in het framework in de DSR. Als de projectkarakteristieken van een project eenmaal gedefinieerd zijn, kan op basis van het framework de juiste projectmethode en processtappen gekozen worden. Bij alle projecten is het belangrijk dat de probleemformulering helder is. Bij routinematige projecten is (in tegenstelling tot eenmalige projecten) automating en support belangrijk voor de ontwikkeling en onderhoud van het model. Conceptualisatie is belangrijk als gewerkt wordt met persoonsgegevens en andere privacy gevoelige informatie. Qua projectmethode worden meerdere iteratiemogelijkheden gebruikt. Tevens kan er gebruikt worden gemaakt van Agile Scrum of Agile Kanban (of eventueel een hybride vorm). Deze keuze is vaak afhankelijk van de ervaring en voorkeur van de projectleider of het projectteam. Agile Scrum is alleen niet passend als processtappen de lengte van een sprint overschrijden.
Het framework is in een casestudy doormiddel van expertsinterviews en focusgroepinterviews getoetst bij een middelgrote onderneming in de tech-sector. Tijdens de focusgroepsessies is dit framework gevalideert door de deelnemers, het dient wel nog in de praktijk getest te worden. Voor generalisatie is een vergelijkbaar onderzoek binnen andere organisaties nog wenselijk.

Summary

Research has shown that a large part of the data science projects were not successful. This is particularly regrettable since data science helps organizations to extract the right knowledge and information from the enormous amount of data that is available today. Follow-up research also found that as many as 82% of the surveyed data sciencists indicated that they did not use the available project methods for data science projects, while 85% indicated that they expected that implementing a better process approach will lead to better project results. This thesis is a study of the influence of the different properties, the project characteristics of a data science project on the choice of the best process steps and/or project methods. A Design Science Research was used for this research. The aim is to create a framework that provides insight and overview of the project characteristics that influence the process steps and thus the project approach of a data science project. This framework can thus contribute to a successful(er) approach to data science projects. The research provides answers to the following research question:

'How can project methods/process steps be used to make a data science project successful?'

The literature review first of all shows an overview of the project methods and the process steps for data science projects. The CRISP-DM model, which is still used in practice, was taken as a basis. This model has been supplemented with a number of process steps: Life cycle selection Problem formulating (as an additional part of the step business understanding), Conceptualization, Automate and Support / Maintenance, plus process themes, namely iteration possibilities and Agile methodologies such as Scrum and Kanban, to reflect the developments of data science project methods. Subsequently, it was investigated which project characteristics for data science projects are distinguished. These are divided into five categories: project types, Data context, Analytical context, Team context and Organizational context.
There is no one-fits-all method for all data science projects, the choice of the right project method and process steps depends on the characteristics of the data science project. This dependency is reflected in the framework in the DSR. Once the project characteristics of a project have been defined, the right project method and process steps can be chosen based on the framework. In all projects, it is important that the problem formulation is clear. In routine projects,(as opposed to one-off projects), automation and support is important for the development and maintenance of the model. Conceptualization is important when working with personal data and other privacy sensitive information. In terms of project method, several iteration possibilities are used. Agile Scrum or Agile Kanban (or possibly a hybrid form) can also be used. This choice often depends on the experience and preference of the project leader or the project team. Agile Scrum is only not preferable if process steps exceed the length of a sprint.
The framework was tested in a case study through expert interviews and focus group interviews at a medium-sized company in the tech sector. During the focus group sessions this framework is caseided by the participants, it still needs to be tested in practice. For generalization, a similar research within other organizations is still desirable.

Inhoudsopgave

1. Introductie

1.1. Aanleiding

De 21e eeuw wordt gezien als de eeuw van de data. Het gebruik van sensoren, machines, apps en software, social media etc. voorziet in een enorme hoeveelheid beschikbare data. Een veel gebruikte term voor deze grote hoeveelheid data is Big Data. De betekenis van Big Data verschilt voor verschillende personen met verschillende achtergronden. Een veel gebruikte definitie is: Data die wordt gekenmerkt door zijn volume, variëteit en snelheid die het bereik van veelgebruikte hardware omgevingen en / of mogelijkheden van actieve softwaretools overschrijdt (Sharda, R., Delen, D., & Turban, E., 2018). Om uit al deze data de juiste kennis en informatie te halen is het vakgebied Knowledge Discovery ontstaan (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). Later (mede door de ontwikkeling) werd voor dit vakgebied ook de termen data analytics of data science gebruikt (Chen, H., Chiang, R. H. L., and Storey, V. C. 2012). Ondanks dat deze termen inhoudelijk niet dezelfde betekenis hebben, zal vanaf nu voor de leesbaarheid van dit rapport de term data science worden gebruikt. Deze kennis en informatie geeft organisaties de mogelijkheid betere beslissingen te nemen en daarmee de juiste acties te ondernemen bij het creëren van concurrentievoordeel (Provost & Fawcett, 2013).

Het onderzoeksgebied waar dit onderzoek plaats vindt is het gebied van Data Science (projecten). Doorgaans wordt een data science project beschreven als een project dat statistische en machine-learning technieken gebruikt om grote hoeveelheden ongestructureerde en / of gestructureerde data, welke zijn gegenereerd door systemen, mensen, sensoren of digitale sporen van informatie van mensen, te extraheren. Dit werk wordt gedaan in een gedistribueerde computeromgeving met als doel correlaties en causale relaties te identificeren, gebeurtenissen te classificeren en te voorspellen, patronen en afwijkingen te identificeren, en waarschijnlijkheden, interesse en sentiment af te leiden (Das, Cui, Campbell, Agrawal, & Ramnath, 2015). Om deze data science projecten effectief te coördineren, zijn in de loop der jaren diverse procesmodellen ontwikkeld. De bekendste hiervan is de CRISP-DM-norm (Cross Industry Standard Process for Data Mining), welke al in de jaren negentig is vastgesteld. CRISP-DM is een voorbeeld van een procesmodel voor datamining voor experts in datamining en biedt een vergelijkbare stapsgewijze procesbeschrijving (Chapman et al., 2000; Wirth, R., & Hipp, J. 2000). Het model noemt zes fasen op hoog niveau: Business understanding, Data understanding, Data preperation, Modelling, Evaluation en Deployment. Later zijn nog meerdere onderzoeken gedaan naar de ontwikkeling van projectmethodieken voor data science projecten (Baijens & Helms, 2019; Li et al., 2016; Mariscal, Marban, & Fernandez, 2010).

In dit rapport wordt de bijdrage van de verschillende projectmethoden aan het resultaat van data science projecten onderzocht en vervolgens getoetst door middel van een case studie.

1.2. Probleemstelling

Het probleem wat onderzocht wordt is waarom data science projecten niet altijd even succesvol zijn. Volgens een onderzoek van Gartner in 2015 zouden namelijk maar liefst 60% van de data science projecten niet worden afgerond en niet verder komen dan de pilot en experiment fase (Saltz, 2015). Saltz (2015), geeft aan dat er processtappen vergeten kunnen worden als niet het juiste proces wordt gevolgd. Saltz, Hotz, Wild, & Stirling (2018) hebben vervolgens onderzoek gedaan in hoeverre het gebruik (of juist het gebrek aan gebruik) van procesmodellen invloed heeft het succes (of juist het gebrek aan succes) van data science projecten. Tijdens dit onderzoek werd gekeken of er in de praktijk bij het uitvoeren van een data science project ook gebruik van een procesmodel werd

gemaakt. 82% van de ondervraagde data sciencisten gaf aan geen procesmodel te gebruiken, terwijl 85% aangaf dat ze verwachten dat het uitvoeren van een betere procesaanpak tot betere resultaten zal leiden.

1.3. Opdrachtformulering

Om een oplossing te vinden op het probleem waarom data science projecten niet altijd even succesvol zijn, wordt in dit rapport onderstaande onderzoeksvraag behandeld. Er is namelijk al onderzoek gedaan naar de ontwikkeling van projectmethoden voor data science projecten (Baijens & Helms, 2019; Li et al., 2016; Mariscal et al, 2010) en het verband van het gebruiken van een projectmodel voor het succes van een data science project. Maar nog niet naar het verband tussen de typen data science projecten en de te nemen stappen (Baijens & Helms, 2019; Li et al., 2016; Saltz et al., 2018).:

Hoe kunnen projectmethoden/processtappen worden gebruikt om een data science project succesvol te maken?

Deze onderzoeksvraag wordt onderverdeeld in de volgende sub-vragen:

- Welke projectmethoden of processtappen voor data science projecten kent de literatuur (literatuurstudie)
- Welke typen data science projecten kent de literatuur/ wat zijn hun karakteristieken? (literatuurstudie)
- Welke projectstappen kunnen worden gebruikt om tot een juiste projectmethode te komen? (literatuur studie/case studie)

1.4. Motivatie / relevantie

Doel van het onderzoek is om een theoretisch kader te scheppen op basis van de beschikbare wetenschappelijke literatuur over data science projecten en de gebruikte projectmethoden. Uit de verschillende wetenschappelijke artikelen is gebleken dat er verschillende projectmethoden toegepast kunnen worden, er is echter nog geen design science research gedaan naar de juiste projectmethode per type project. Hier zal dit rapport aan bijdragen, de opgedane kennis in dit rapport kan worden gebruikt voor verder onderzoek op dit gebied. Allereerst wordt door een literatuurstudie een theoretisch overzicht gecreëerd waarin de deelvragen worden behandeld. Dit overzicht of raamwerk laat zien welke projectmethoden per type data science project gebruikt kunnen worden. Vervolgens wordt dit theoretisch overzicht in de praktijk getoetst binnen een case organisatie waar verschillende data science projecten uitgevoerd worden. Ten slotte zullen de resultaten van dit onderzoek de organisatie helpen met de keuze van de juiste projectmethoden voor hun toekomstige data science projecten

1.5. Aanpak in hoofdlijnen

Om de onderzoeksvragen te kunnen beantwoorden is gekozen voor de Design Science Researce Methodology (DSRM). Peffers, Tuunanen, Rothenberger, & Chatterjee. (2007) heeft een model geïntroduceerd met daarin de stappen voor een DSRM.

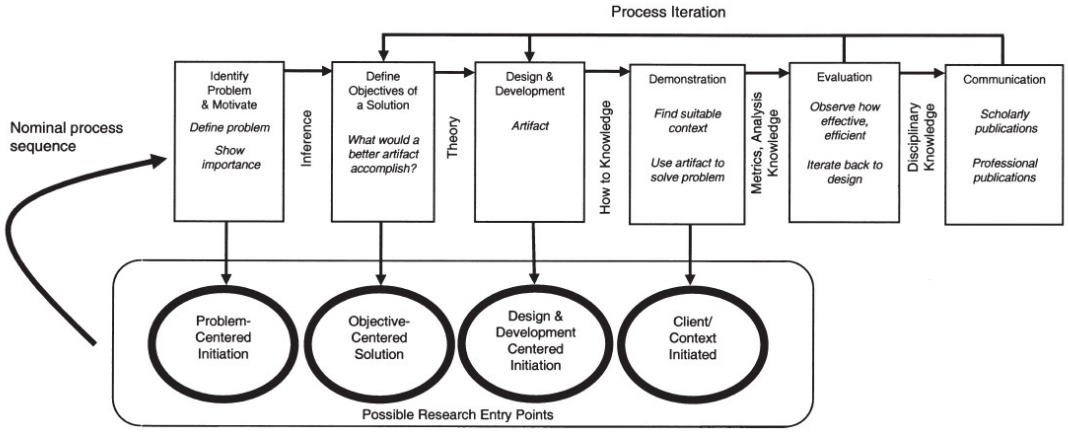


Figure 1 DSRM Process Model

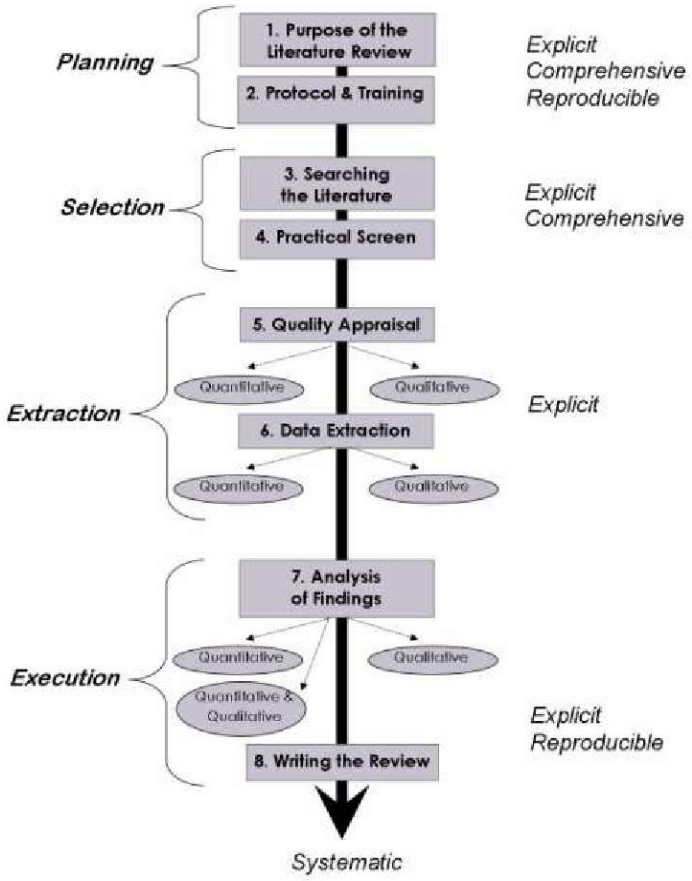
In hoofdstuk 1 begint het onderzoek met het identificeren en definiëren van een probleemstelling en onderzoeksvragen. Hierbij is ook de relevantie voor zowel de wetenschap als de praktische relevantie beschreven. In hoofdstuk 2 is vervolgens een theoretisch onderzoek gedaan welke is samengevat in een theoretisch framework. Dit onderzoek bevat de huidige theorie rondom de projectaanpak voor data analyse projecten. Doel van dit onderzoek is om de onderzoeksvragen te beantwoorden. In hoofdstuk 3 staat een beschrijving en verantwoording van de gevolgde onderzoeksmethode. In hoofdstuk 4 is het framework vervolgens getoetst door middel van een case study binnen een jonge technische onderneming in Nederland welke opereert in een niche markt en zijn producten wereldwijd afzet. Hiervoor zijn binnen de organisatie diverse onderzoeksmethoden toegepast zoals interviews, groepssessies / focusgroepen en desk research naar bestaande data science projecten. In hoofdstuk 5 staan tenslotte de discussie, conclusies en aanbevelingen voor de praktijk en vervolgonderzoek. Uiteindelijk zijn de uitkomsten van het literatuuronderzoek en de toetsing binnen de organisatie samengevat en gepresenteerd. Doel van deze presentatie is om aan te tonen of de theoretische oplossing ook in de praktijk uitgevoerd kan worden.

2. Theoretisch kader

Dit hoofdstuk behandeld de bestaande wetenschappelijke literatuur over het onderzoeksprobleem. Dit hoofdstuk presenteert de aanpak, implementatie en resultaten van de literatuurstudie. In deze literatuurstudie worden de deelvragen over de processtappen en de projectkarakteristieken beantwoord.

2.1. Onderzoeksaanpak

Als handleiding voor de literatuurstudie is het model van Okoli and Schabram (2011) gebruikt. Zij behandelen een concreet stappenplan, Systematic Literature Review (SLR) (figuur 2), voor het uitvoeren van een literatuurstudie. In dit onderzoek worden deze stappen doorlopen om op wetenschappelijk verantwoorde manier een theoretisch kader te vormen rondom het onderwerp van deze scriptie. De uitvoering hiervan wordt in de volgende paragraaf uitgewerkt).



Figuur 2 Systematic Literature Review (Okoli & Schabram, 2010)

Deze literatuurstudie probeert antwoord te geven op de onderzoeksvraag en sub-vragen uit hoofdstuk 1.3. De uitkomsten van deze literatuurstudie zullen worden weergegeven in een overzicht om de ontwikkeling van het onderzoeksgebied weer te geven.

Bovenstaande uitleg van de onderzoeksaanpak is een samenvatting. De gedetailleerdere versie is terug te vinden in bijlage 1

2.2. Uitvoering

Okoli and Schabram (2010) benadrukken het belang vooraf een protocol voor de literatuurstudie vast te stellen. Hiervoor is het Systematic Literature Review (SLR) van Okoli and Schabram (2011) gebruikt.

Als startset is gebruik gemaakt van zes artikelen welke door de begeleiders van de Open Universiteit beschikbaar waren gesteld. Deze zijn gebruikt voor de gebiedsverkenning en voor het opstellen van de juiste zoektermen voor de query. Als query is uiteindelijk gekozen voor: TI= ((Data analytics OR Data Science OR knowledge discovery) AND (Project* OR Process)). In eerste instantie is gezocht vanuit de bibliotheek van de Open Universiteit, echter was het daar niet mogelijk de resultaten in 1 overzicht te exporteren voor analyse. Als zoekmachines is in overleg met de begeleider gekozen voor de databases van Web of Science en IEEE in verband met hun affiniteit met het onderwerp. Deze databases werden benaderd vanuit de bibliotheek van de Open Universiteit.

Deze query leverde bij Web of Science 47 unieke artikelen en IEEE 96 unieke artikelen op. Na de practical screening (op basis van titel, samenvatting, inleiding, conclusie) bleven er 12 artikelen over, na snowballing kwamen hier nog 10 artikelen bij. Dit resulteerde in een uiteindelijke set van 22 artikelen voor de literatuurstudie.

Bovenstaande uitvoering is een samenvatting. De gedetailleerdere uitvoering is terug te vinden in bijlage 1

2.3. Resultaten en conclusies

2.3.1. Projectmethoden/Procesmodellen

Het oudst bekende procesmodel voor het analyseren van data is het Knowledge discovery in databases (KDD) model (Fayyad, Piatetsky-Shapiro, and Smyth, 1996) als voorloper op data science. In 2000 werd het CRISP-DM model met 6 processtappen ontworpen en gepubliceerd (Chapman et al., 2000) . Het CRISP-DM model geldt nog steeds als het bekendste procesmodel binnen de data-science en wordt gezien als de standaard door veel data analysten. In 2010 hebben Mariscal et al. de ontwikkeling van diverse Knowledge discovery en Data Mining schematisch weergegeven en de stappen hieruit in het Refined Data Mining Process samengevat in een nieuw uitgebreider proces van 17 processtappen. Baijens et al. hebben in 2019 een overzicht gemaakt van de aanvullingen op het proces van Mariscal (2010) en in vergelijken deze ook met het CRISP-DM model.

CRISP-DM

Het meest bekende procesmodel voor data science projecten is het CRISP-DM model (Chapman et al., 2000). Dit procesmodel model bevat zes processtappen.

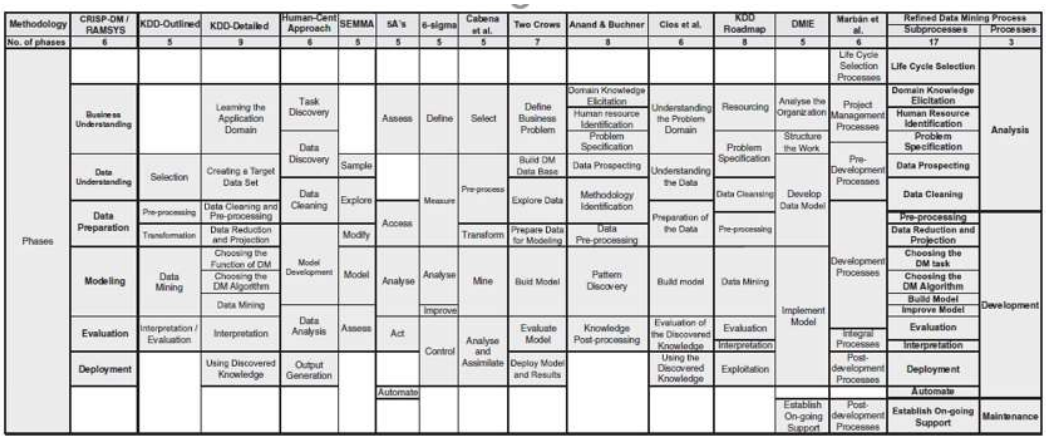
- Business understanding – Bepaal de business doelstellingen, doe een omgevingsanalyse, bepaal de data science doelstelling, ontwikkel een projectplan.
- Data Understanding – Verzamel initiële data, beschrijf deze data, onderzoek deze data, verifieer de kwaliteit van deze data.
- Data Preperation – Selecteer de data, schoon deze data op, (re)construeer deze data, integreer deze data, format deze data.
- Modeling – Selecteer een modeling techniek, maak een test ontwerp, bouw het model, beoordeel/evalueer het model.
- Evaluation – Evalueer de resultaten, evalueer het proces, bepaal de vervolgstappen

- Deployment – Plan de uitrol van het model, plan monitoring en onderhoud van het model, rapporteer het uiteindelijke resultaat van het model en het proces, review het project.

Oorspronkelijk was het CRISP-DM model ontworpen met iteratieve stappen, maar werd in de praktijk als waterval toegepast (Mariscal et al., 2010) Het model mist een aantal projectmanagement activiteiten (Mariscal et al., 2010) en houdt helaas geen rekening met actuele procesmanagement activiteiten zoals kwaliteitsmanagement en verandermanagement (Li et al., 2016).

Toegevoegde stappen

Later zijn nog meerdere procesmodellen ontwikkeld. Sommige van deze modellen hebben stappen toegevoegd aan CRISP-DM. Anderen hebben een verdieping toegevoegd aan CRISP-DM. Tenslotte zijn er methoden ontwikkeld met een flexibeler proces dan de (gebruikte) watervalmethode van CRISP-DM. Mariscal et al. hebben in 2010 een vergelijking gemaakt waarin zij 14 procesmodellen met elkaar hebben vergeleken. Als basis is hiervoor het CRISP-DM procesmodel genomen. In de vergelijking is gekeken wat de verschillen en de aanvullingen ten opzichte van het CRISP-DM procesmodel zijn. Vervolgens hebben de bevindingen van deze vergelijking geresulteerd in een her-gedefinieerd procesmodel welke bestaat uit drie hoofdfasen en in totaal 17 processtappen, zie figuur 3.



Figuur 3 Overzicht Refined Data Mining Process (Mariscal et al. 2010)

Belangrijkste reden om het CRISP-DM model aan te passen is dat dit model een aantal project management activiteiten mist (Mariscal et al., 2010). Het her-gedefinieerde procesmodel bevat met 17 processtappen veel meer stappen dan het CRISP-DM procesmodel en ook als de andere modellen uit de vergelijking. De reden dat hiervoor is gekozen is omdat de onderlinge afhankelijkheden tussen de processtappen zo beter weergegeven kunnen worden (Mariscal et al., 2010). Als er gekeken wordt naar de inhoudelijke toevoeging, zijn alleen de eerste stap (Life Cycle Selection) en de laatste twee stappen (Automate en Establish On-going Support), een toevoeging ten opzichte van CRISP-DM. De overige stappen zijn een verdieping of opsplitsing van één van de CRISP-DM stappen.

- Life Cycle Selection - Identificeren en selecteren van acquisitie, levering en levenscyclus op basis van het type project. CRISP-DM omvat helemaal geen acquisitie- of leveringsprocessen, deze zijn wel noodzakelijk (Mariscal et al., 2010). Hier wordt ook bepaald welke projectmanagementmethode gebruikt wordt. Reden hiervoor is dat voor de iteratieve stappen een meer agile of hybride methodologie vereist is in plaats van de meer traditionele watervalbenadering van CRISP-DM (Baijens and Helms, 2019)

- Automate – Automatiseren van het datamining-proces om met name datamininggebruikers die geen expert zijn, in staat te stellen eerder verkregen modellen toe te passen op nieuwe data. (Mariscal et al., 2010).
- Establish On-going Support - Onderhoud en support voor het model zoals backups, onderhoud, model updates en software updates. Deze stap is noodzakelijk als de uitkomst het data science project vaker dan éénmalig gebruikt zal worden (Mariscal et al., 2010).

Baijens & Helms hebben in 2019 een overzicht gepubliceerd van de procesmodellen welke waren gepubliceerd na het onderzoek van Mariscal (2010)(figuur 4), met daarin de vergelijk tussen deze modellen en het CRISP-DM model en van het model van Mariscal (2010). Zij onderzochten daarin of de extra genoemde stappen daadwerkelijk toevoegingen waren, of dat deze eigenlijk al als onderdeel in één van de andere stappen zat.

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

Figuur 4 Overzicht procesmodellen Baijens & Helms (2019)

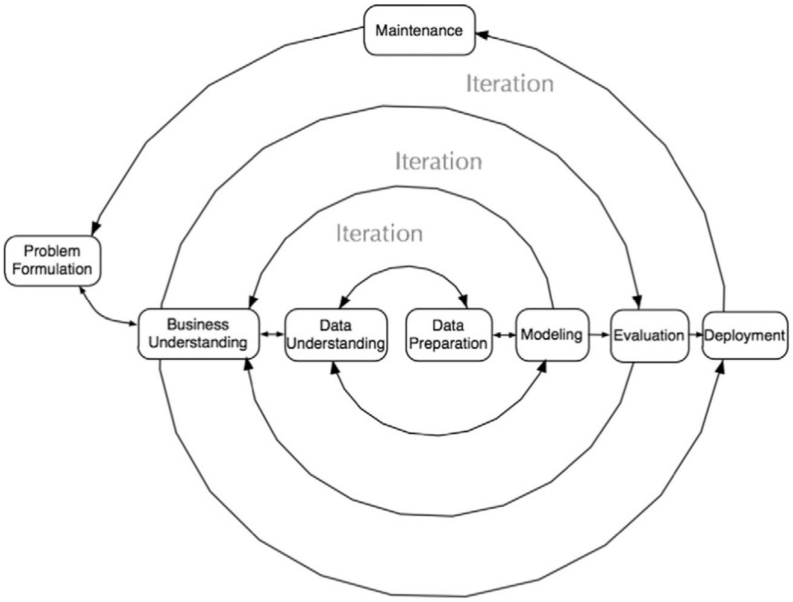
Van de procesmodellen welke na het procesmodel van Mariscal (2010) waren ontstaan, heeft alleen het model van Ahangama en Poo (2015) nog een aanvullende stap ten opzichte van het model van Mariscal (2010), namelijk de stap Conceptualization (Baijens & Helms, 2019). Deze stap hebben zij ingevoegd tussen de stappen Data Understanding en Data Preperation. Doel van deze conceptualisatiestap is het verkennen van variabelen die zullen worden gebruikt en de relaties tussen die variabelen (Ahangama en Poo 2015). Deze stap bestaat uit een literatuuronderzoek naar het domeindoel, formuleren van een onderzoeksvraag en de ontwikkeling van een conceptueel model met beschrijving van de variabelen die in het model worden gebruikt. De reden dat deze stap

is toegevoegd, is dat de analysetechniek niet afhankelijk moet zijn van de beschikbare gegevens, maar van het doel van de organisatie (Ahangama en Poo 2015). Oorspronkelijk is deze stap toegevoegd voor gebruik in de gezondheidszorg waar beslissingen effect kunnen hebben op de zorg voor de patiënt, maar deze stap kan worden gegeneraliseerd (Baijens & Helms, 2019).

Vergeleken met CRISP-DM hebben (Li, Thomas, & Osei-Bryson, 2016) aan hun snail shell model twee stappen toegevoegd: Problem formulating (als aanvullende stap na Business Understanding) en vergelijkbaar aan de stap Problem Specification van Mariscal (2010) en een laatste stap Maintenance, te vergelijken met de laatste stap van Mariscal (2010), Establish On-going Support. Ook Ahangama en Poo (2015) en Grady (2016) geven hun invulling aan deze Maintenance stap.

Aangepast proces

Naast deze aanvullende stappen hebben diverse auteurs ook ruimte in hun projectmethode voor een aangepast/verbeterd proces. Li et al. (2016) heeft bijvoorbeeld een aangepast data science model ontwikkeld, met daarin 8 processtappen en met name de iteratieve mogelijkheden tussen de processtappen, zie figuur 5. Deze iteratieve mogelijkheden dragen bij aan een agile manier van werken omdat zij gedurende het project de mogelijkheid voor terugkoppeling, evaluatie en daardoor optimalisatie mogelijk maken. Door deze agile methoden toe te passen op een watervalproces ontstaat dus een hybride procesmodel.



Figuur 5 Snail shell model (Li et al., 2016)

Verder wordt er in het Snail shell model, naast dat er wordt aangegeven welke stappen er genomen moeten worden (wat er moet gebeuren), ook aangegeven hoe deze stappen genomen moeten worden (hoe deze moeten gebeuren).

Saltz (2017) schrijft over een complete toepassing van Agile methodieken als Scrum en Kanban voor data science projecten:

- Agile Scrum – Vooraf worden de teamleden geïnstrueerd te werken in korte perioden (sprints) van 2-4 weken. Als eerste stap van elke sprint worden de doelen voor de komende sprint bepaald op basis van haalbaarheid en prioriteit. Gedurende de sprint wordt dan ook alleen aan deze zaken gewerkt. Eventuele nieuwe zaken worden tijdens een volgende sprint

opgepakt. Aan het einde van de sprint worden de vooraf gestelde doelen ook opgeleverd (Saltz, 2017). Een voordeel van deze methode is dat er na elke sprint een levering plaats vindt (en er dus continue voortgang zichtbaar is). Een nadeel van deze methode is dat het voorafgaand van aan project lastig is het definitieve tijdlijn en ook deliverables aan te geven (beide kunnen namelijk gedurende het project worden bijgesteld)

- Agile Kanban – Kanban is van pipeline procesmethode oorsprong bedacht voor lean manufacturing. In combinatie met een aantal processtappen uit de bekende procesmethoden voor data science projecten en dit ook voor data science projecten toe te passen. Belangrijke tool is het Kanban bord waar lopende zaken kunnen worden beheert. Deze lopende zaken worden ingedeeld in categorieën als nieuw, voorbereiding, analyse, implementatie, afgerond. Vooraf wordt bepaald hoeveel zaken in elke categorie toegestaan zijn, hierdoor moet dus worden geprioriteerd. Prioritering vindt plaats op basis van hoog over user-stories (Saltz, 2017). Een voordeel van deze methode is dat er relatief weinig work in progress is (door maximaal aantal zaken per fase)

Samenvatting - Aanvullingen ten opzichte van CRISP-DM

Als bovenstaande projectmethoden en procesmodellen worden vergeleken met CRISP-DM zijn er een aantal aanvullingen.

Allereerst zijn er toegevoegde stappen/thema's:

- Life cycle selection (Mariscal et al., 2010). De inhoud van deze stap wordt helemaal niet meegenomen in CRISP-DM. In het Snail shell model wordt de inhoud van deze stap opgenomen in de beginstap Business understanding. De inhoudelijke toevoeging is dat er vooraf processen voor acquisitie, levering en levenscyclus worden vastgesteld. Hier wordt ook de projectmanagementmethode bepaald.
- Conceptualization (Ahangama and Poo, 2016). De inhoudelijke toevoeging is dat de analysetechniek niet afhankelijk moet zijn van de beschikbare gegevens, maar van het doel van de organisatie.
- Automate (Mariscal et al., 2010). De inhoudelijke toevoeging van deze stap is dat het datamining-proces wordt geautomatiseerd voor hergebruik. Deze stap is met name handig voor projecten waarvan de verwachting is dat het gebruikte proces vaker gebruikt zal worden en ook beschikbaar moet zijn voor niet-expert.
- Support / Maintanance (Mariscal et al., 2010; Ahangama and Poo, 2015; Lie et al., 2016). De inhoudelijke toevoeging van deze stap is dat er processen voor backups, onderhoud, model updates en software updates worden toegevoegd, plus eventuele monitoring van het model. Deze stap is noodzakelijk als de uitkomst het data science project vaker dan éénmalig gebruikt zal worden.

Daarnaast de benadering van het proces.

- Hybride procesbenadering. CRISP-DM wordt in de praktijk toegepast als een waterval gedreven procesmodel. Hier volgen de stappen elkaar op en is er bijna geen mogelijkheid nog een stap terug te gaan. Dit is een nadeel als er tijdens het proces gewenste aanpassingen zijn. Voordeel kan zijn dat als vooraf precies bekend is wat het resultaat moet worden, dit beter planbaar is. Het model van Mariscal et al. (2010) en het Snail shell zijn al meer hybride procesmethoden. Deze bevatten iteratieve stappen waarin tussentijds geleverd wordt en ook tussentijds bijgestuurd kan worden. Indien gewenst kan een stap worden herhaald, of de cyclus opnieuw worden doorlopen.

- Agile methoden zoals Scrum en Kanban. Saltz (2017) laat zien hoe Agile methodieken als Scrum en Kanban kunnen worden toegepast op data science projecten. Hier worden continue in korte perioden zaken opgeleverd en zo gebouwd aan een oplossing. Voordeel van deze methoden is dat zij vaak beter aansluiten op de behoeften van de gebruiker aangezien continue bijgestuurd en geprioriteerd wordt. Nadeel is dat het (onder andere budgettair) lastiger planbaar is.

2.3.2. Kenmerken van data science projecten

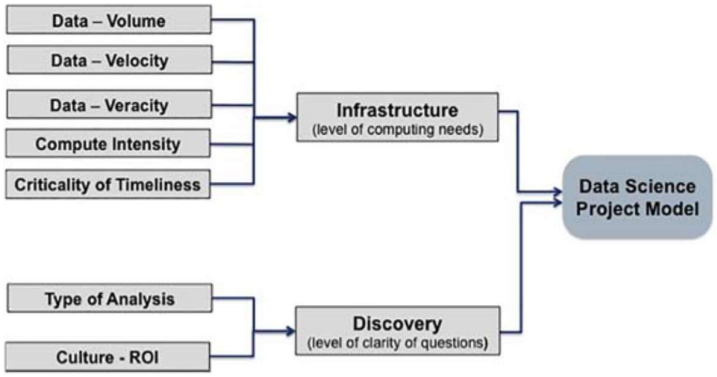
Er wordt door Saltz en Shamshurin (2015) onderscheid gemaakt tussen twee typen data science projecten, routinematige en eenmalige projecten. Bij de eenmalige projecten worden meestal ad-hoc vragen of problemen onderzocht. Deze projecten zijn vaak minder gestructureerd dan routinematige projecten (Saltz & Shamshurin, 2015). Bij de routinematige projecten worden meestal modellen of rapporten gemaakt om een bedrijfsproces te ondersteunen, (Ahangama & Poo, 2015; Li et al., 2016).

Volgens Saltz (2017) zijn er 14 kenmerken voor data science projecten welke hij vervolgens samenvat in vier hoofdkenmerken: Data Context, Analytical Context, Team Context en Organizational Context. Als een data science project vooraf wordt geanalyseerd op basis van het framework met deze 14 kenmerken, wordt beter in kaart gebracht waar de mogelijke project uitdagingen zijn en geeft dit ook beter beeld hoe het project het best kan worden gemanaged.

- Data Context : Variety, Volume, Velocity, Veracity
- Analytical Context: Type of Analysis, Compute Intensity, Criticality of Timelines.
- Team Context: Size of Team, Virtuality of Team, Manager Experience.
- Organizational Context: Size of Organization, Organization Culture & Process, Organization Culture & ROI, Total number of Data Science Teams in Organization.

Bepaalde attributen blijken een grotere impact te hebben op de uitdagingen welke data science teams ondervonden dan anderen en er zijn verbanden tussen sommige attributen. Dit resulteerde in een vereenvoudigd model met twee hoofdkenmerken en zeven onderliggende kenmerken, zie figuur 6 (Saltz, 2017):

- Discovery – Welke vragen worden er binnen het project aangepakt (bijvoorbeeld een duidelijke onderzoeksvraag of juist een algemener onderzoek om waarde uit de data te halen)
    - Type of Analysis, wat is de focus van de analyse: genereren van hypotheses / hypothese testen / beantwoorden van een reeds bekende vraag
    - Organization Culture & ROI, is er binnen de organisatie focus op ROI
- Infrastructure - zijn er aanzienlijke IT-middelen vereist?
    - Volume, grootte/ hoeveelheid van de te analyseren data (TB, GB)
    - Velocity, snelheid van data collectie (Hoog, middel, laag)
    - Veracity, hoe schoon en betrouwbaar is de data (Schoon, vuil)
    - Compute Intensity, rekenintensiviteit van de datavoorbereiding en de modeluitvoering
    - Criticality of Timelines, moet de analyse binnen een specifieke tijd zijn uitgevoerd



Figuur 6 Model for defining a data science project (Saltz, 2017)

Niet alle attributen hebben invloed op de keuze voor een projectmodel. Daarom wordt nu per groep de invloed op de projectmethode/processtappen besproken

Project context

Routinematige projecten. Mariscal et al., (2010) geven aan dat maintenance een belangrijke stap is bij routinematige projecten Eenmalige projecten zijn lastiger voorspelbaar en planbaar, daardoor het lastiger is vooraf een planning en kosten/baten analyse te maken.

Data context

Van de vier kenmerken Variety, Volume, Velocity en Veracity, hebben individueel, Velocity en Veracity weinig tot geen invloed, alleen Variety en Volume is, als de variatie of volume groter is dan in één sprintlengte behandeld kan worden, van invloed op Agile Scrum. Agile scrum is dan minder geschikt. Als data aan al deze kenmerken voldoet, spreken we van Big data. In het geval van Big Data is een iteratief proces gewenst, onder andere in verband met snelle veranderingen (Li et al., 2016)

Analytical context:

Type of Analysis, Compute Intensity en Criticality of Timelines hebben alleen Type of Analysis en Criticality of Timelines invloed. Type of Analysis: Verschillende typen vragen/problemen welke beantwoord dienen te worden, behoeven soms verschillende typen analyse, bijvoorbeeld open analyses (genereren van hypotheses) of gerichte analyses (testen van hypotheses) (Das et al., 2015; Saltz, 2015). Verschillende typen analyse kunnen verschillende manieren van benaderen en daarmee verschillende eisen aan de projectmethode vragen. Een structurele analyse (zoals bijvoorbeeld een BI-dashboard) behoeft een maintenance stap. Voor het onderzoeken van deze context is een goede business understanding en problem formulating nodig. Criticality of Timelines - Als de analyse binnen een specifieke tijd uitgevoerd moet worden, zal gekozen worden voor een projectmethode die hieraan kan voldoen. Een projectmethode volgens een watervalmethode duurt over het algemeen langer dan een projectmethode met parallelle stappen.

Team context

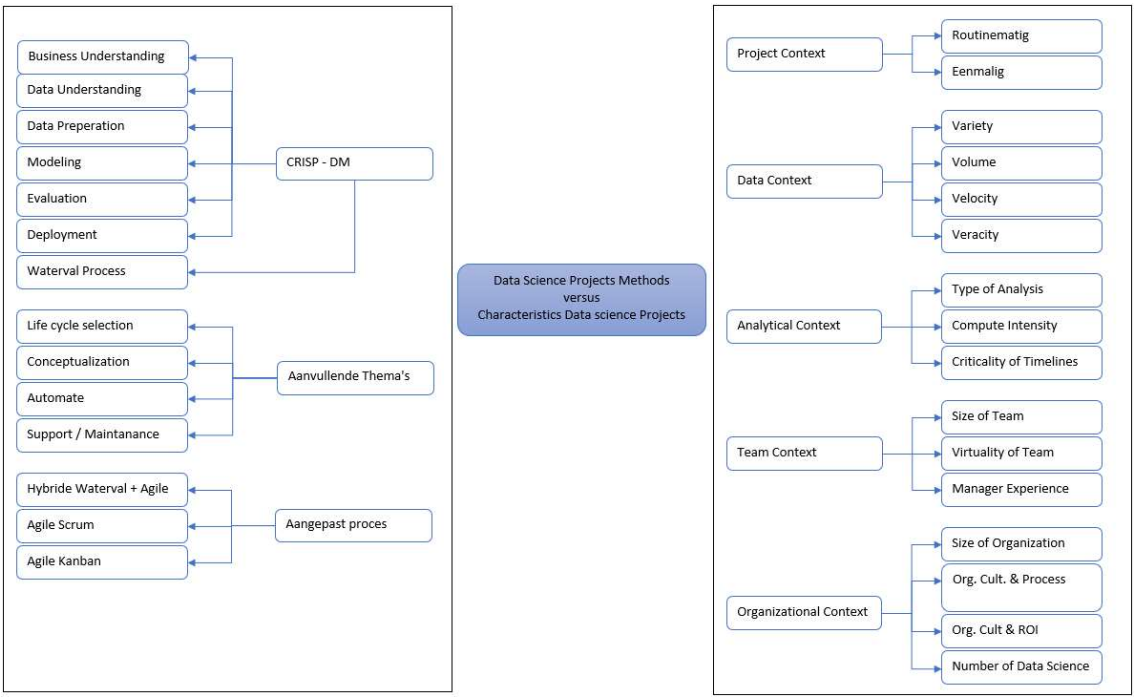
Van de drie kenmerken Size of Team, Virtuality of Team en Manager Experience, hebben Virtuality of Team en met name Manager Experience invloed, Size of Team minder. Virtuality of Team: De samenstelling van het team en daarmee ook de ervaring van de teamleden kunnen invloed hebben op de keuze voor de project methode. Gao et al., (2015) geven aan dat een multidisciplinair team gewenst is. Manager Experience: De achtergrond en de ervaring van de (project)manager zal mede bepalen voor welke projectmethode hij zal kiezen.

Organizational context

Van de vier kenmerken van Size of Organization, Organization Culture & Process, Organization Culture & ROI en Total number of Data Science Teams in Organization, hebben alleen Organization Culture & Process en Organization Culture & ROI invloed. Organization Culture & Process: Als er binnen de organisatie focus ligt op processen, zal hier bij de keuze van de projectmethode rekening worden gehouden. Organization Culture & ROI: Als er binnen de organisatie focus ligt op de ROI, zal hier bij de keuze van de projectmethode rekening worden gehouden. Met name bij lastig planbare projecten, zoals bijvoorbeeld hypothese genererende projecten, is dit lastig. Erg flexibele projectvormen zijn dan minder gewenst.

2.3.3. Theoretisch framework

Naar aanleiding van bovenstaand literatuuronderzoek zien we dat er twee groepen zijn welke invloed hebben op dat science projecten. Enerzijds de data science projects methods, de processtappen en thema's/processen welke worden gebruikt en anderzijds de karakteristieken van de data science projecten. Deze zijn schematisch weergegeven in het theoretisch framework, het Artefact (figuur 7). Dit framework wordt in het vervolgonderzoek in de case study getoetst.



Figuur 7 Theoretisch Framework

2.4. Doel van het vervolgonderzoek

Er zijn diverse projectmethoden voor data science projecten beschikbaar, maar deze worden nog niet (op de juiste manier) toegepast. Daarnaast zijn er ook verschillende typen data science projecten met verschillende eigenschappen, die daardoor een verschillende aanpak en dus verschillende projectmethode behoeven. Het is voor organisaties nog onduidelijk welke projectmethoden het best per type data science project gebruikt kan worden. Deze DSR is daarom bedoeld om een raamwerk te bouwen welke kan worden gebruikt om de juiste projectmethode per type data science project te kiezen.

3. Methodologie

Dit hoofdstuk beschrijft de methodieken die zijn gebruikt tijdens dit onderzoek. Het onderzoek is uitgevoerd volgens de Design Science Research Methodology (DSRM) van Peffers et al. (2007). Als reflectie van de rigor en relevantie is het Design Science Framework van Hevner et al. (2004) toegepast.

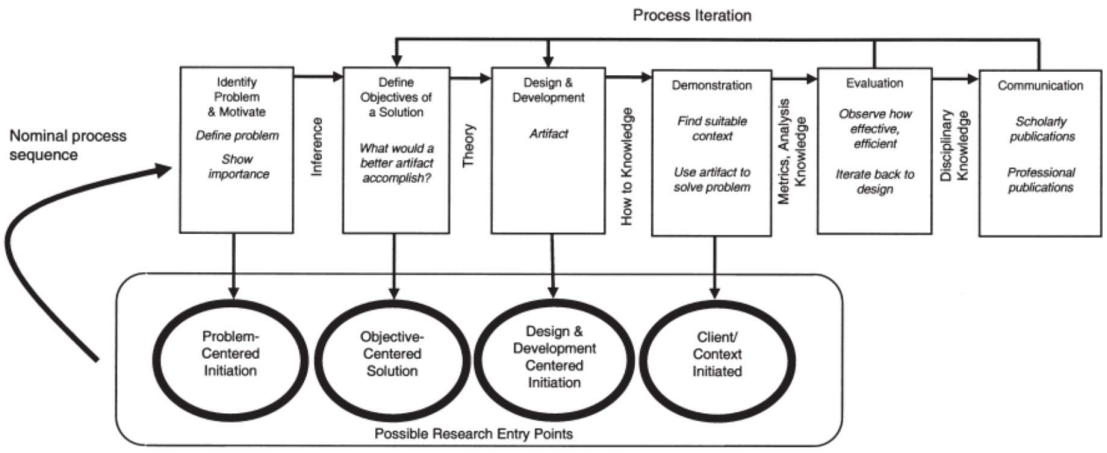
3.1. Conceptueel ontwerp: keuze van onderzoeksmethode(n)

Het doel van dit onderzoek is een artefact te ontwerpen rondom het onderwerp, welke zowel bruikbaar is in de praktijk, als bijdraagt aan de wetenschap. Als onderzoeksmethode is daarom een design science research (DSR) uitgevoerd. Design science research is een ontwerpwetenschap die IT-artefacten creëert en evalueert welke zijn bedoeld om geïdentificeerde organisatorische problemen op te lossen (Peffers, et al., 2007). Deze methode is geschikt voor dit onderzoek omdat er een theoretisch framework (artefact) op basis van wetenschappelijke literatuur wordt ontwikkeld, om inzicht te geven in de projectmethoden voor data science projecten. Bijkomend voordeel van deze onderzoeksmethode is daarnaast ook dat er op een praktische manier invulling wordt gegeven aan het vinden van een juist procesmodel per type data science project. Vanuit het ontworpen framework/artefact wordt een theoretisch antwoord gegeven op de onderzoeksvraag, welke vervolgens door middel van een case study bij een case organisatie in de praktijk is getoetst. Binnen deze case organisatie is doormiddel van interviews, focusgroepsinterviews en desk research de benodigde informatie ingewonnen om het theoretische framework te toetsen.

3.2. Technisch ontwerp: uitwerking van de methode en gegevensanalyse

Voor het ontwerp van een design science research heeft Peffers et al. (2007) 6 processtappen gedefinieerd (figuur 8):

1. **Problem definition**: In de eerste processtap wordt de probleemstelling en de onderzoeksvragen gedefinieerd. In deze stap wordt ook de bijdrage van het onderzoek voor zowel de wetenschap, als ook het bedrijfsleven besproken.
2. **Define the objectives for a solution**: In de tweede processtap worden de doelstellingen van de oplossing besproken.
3. **Design and development**: In de derde processtap wordt het artefact gecreëerd.
4. **Demonstration**: In de vierde processtap wordt het ontworpen artefact getoetst.
5. **Evaluation**: In de vijfde processtap wordt het artefact en de bijdrage aan het probleem geëvalueerd.
6. **Communication**: In de zesde processtap zal het onderzoek worden gecommuniceerd / gepresenteerd aan de stakeholders van het onderzoek, zowel binnen de case-organisatie als binnen de Open Universiteit.



Figuur 8 DSRM Process Model (Peffers et al., 2007)

3.2.1. Problem definition

Het probleem wat onderzocht wordt, staat in hoofdstuk 1 van dit onderzoek. Data science projecten zijn niet altijd even succesvol. 60% van de data science projecten niet worden afgerond en niet verder komen dan de pilot en experiment fase. Het gebruik van de juiste projectmethoden zou een positief effect op het succes van data kunnen hebben.

3.2.2. Define the objectives for a solution

Doelstelling van dit onderzoek betreft het een kwalitatieve oplossing, waarin een artefact wordt ontwikkeld om de probleemstelling te beantwoorden. Doelstelling van dit artefact is een goed overzicht te geven waarin de verschillende eigenschappen van data science projecten worden afgezet tegen de verschillende mogelijke processtappen en projectmethoden van data science projecten.

3.2.3. Design and development

In de derde processtap wordt het artefact gecreëerd over de projectmethoden voor data science projecten. Hiervoor is gestart met een literatuuronderzoek voor het creëren van een theoretisch kader ter beantwoording van de deelvragen. Vervolgens is op basis van dit theoretisch kader het artefact gecreëerd. Dit artefact is een systematisch model waarin verschillende eigenschappen van data science projecten worden afgezet tegen de verschillende mogelijke processtappen en vormen van data science projecten.

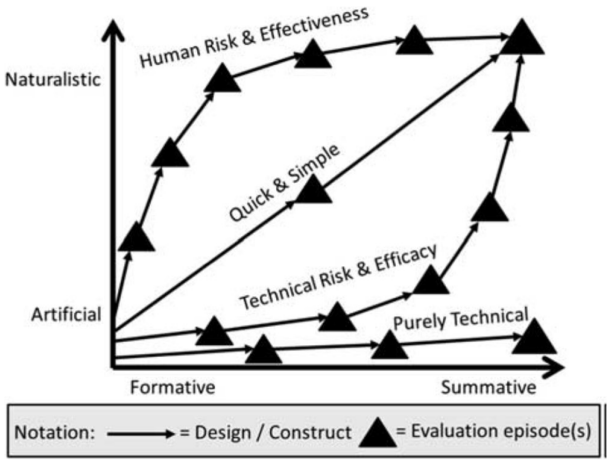
3.2.4. Demonstration

In de demonstratie wordt het ontworpen artefact doormiddel van een casestudie getoetst bij een case organisatie. Door een casestudie in een echte organisatie uit te voeren, worden complexiteiten uit een echte organisatie meegenomen (Venable et al., 2016).
Binnen deze case organisatie worden eerst semigestructureerde interviews afgenomen bij diverse stakeholders van data science projecten. Voordeel van semigestructureerde interviews is dat er naast de voorbereide vragen ook verdiepende vragen over besproken thema's gesteld kunnen worden (Saunders Lewis, & Thornhill. 2016). Van deze stakeholders wordt verwacht dat zij ervaring hebben met het uitvoeren van data science projecten en inzicht hebben in de projectmethode welke is gevolgd bij deze projecten. Tijdens deze interviews wordt inzicht gecreëerd over gebruikte

projectstappen en methoden, plus inzicht in de eigenschappen van deze projecten. Vervolgens zal er deskresearch worden verricht naar de huidige of recente data science projecten. Hierna vindt nog een verdiepende interviewronde plaats door middel van een focusgroep interview. Het voordeel van een focusgroep interview is dat de geinterviewden samen kunnen discussiëren over de vragen en antwoorden. Hierdoor kunnen ze ook elkaar bevragen en daarmee tot een gezamenlijk volledig(er) antwoord komen (Saunders, et al., 2016).
Door meerdere onderzoeksmethoden en daardoor meerdere bronnen te gebruiken wordt door triangulatie de validiteit verbeterd.

3.2.5. Evaluation

Na de demonstratie wordt het model geëvalueerd. De evaluatie binnen de caseorganisatie gebeurt door middel van een tweede focusgroep interview met diverse stakeholders van data science projecten. Na deze evaluatie wordt het model eventueel aangepast. Venable et al. (2016) heeft voor de evaluatie een framework for evaluation in design science (FEDS) ontworpen. In onderstaand figuur 9 is deze schematisch weergegeven met twee dimensies welke beide worden geëvalueerd.

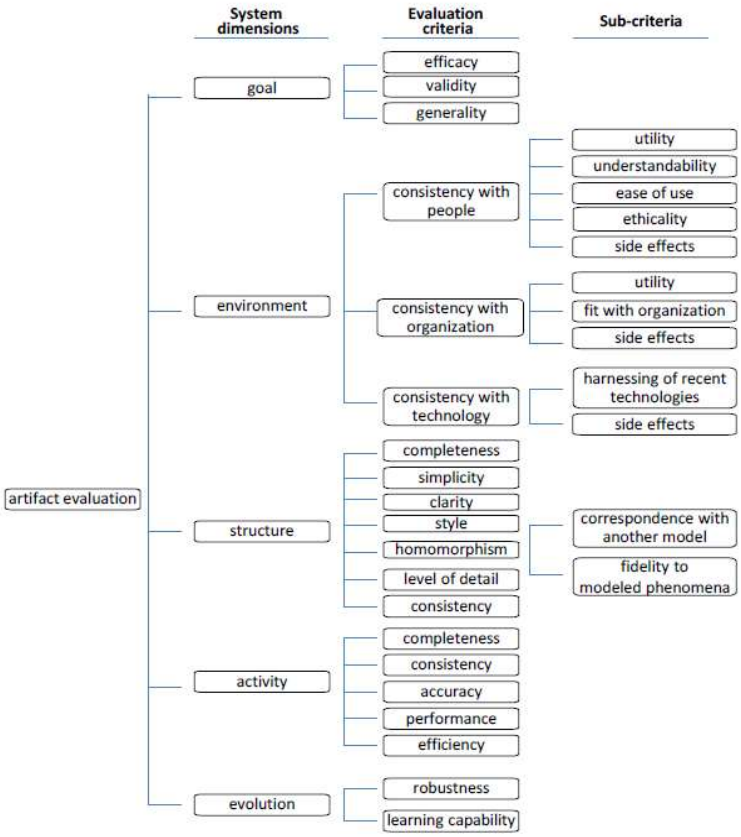


Figuur 9 Framework for evaluation in design science (FEDS)

Naturalistic evaluation onderzoekt de oplossing binnen een echte omgeving, in dit geval een case organisatie. Door deze in een echte organisatie uit te voeren, worden alle complexiteiten uit een echte organisatie meegenomen (Venable et al., 2016) en onderzoeken we de praktische relevantie. Volgens het FEDS framework is dit een 'Human Risk & Effectiveness' evaluatiestrategie. De Formative evaluation is bedoeld om het artefact verder te verbeteren. In dit rapport gebeurt dat door middel van interviews met experts, deskresearch en een validatiepresentatie tijdens de ontwerp- en ontwikkelingsfase. De summative evaluation is bedoeld om het artefact te valideren.

Prat, Comyn-Wattiau, en Akoka. (2014) deden onderzoek naar de evaluatie van DSR artefacten. Zij vermelden dat design science researcers vaak worstelen met de evaluatie van het artefact. Zij hebben vervolgens een framework met evaluatiecriteria ontworpen gebaseerd op de vijf belangrijkste System dimensions voor het artefact (goal, environment, structure, activity, and

evolution), om het artefact te evalueren, zie figuur 10. Dit model wordt gebruikt voor de summatieve evalutie.



Figuur 10 Evaluatiecriteria voor Artifact evaluation (Prat et al., 2014).

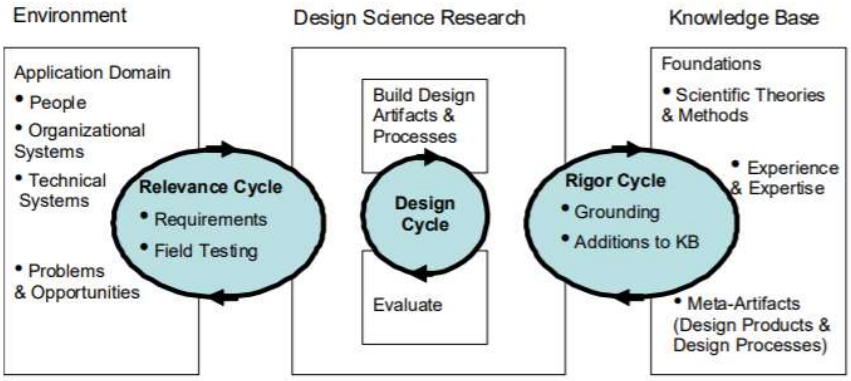
3.2.6. Communication

Tenslotte zal het onderzoek zal worden gecommuniceerd/gepresenteerd aan de stakeholders van het onderzoek, zowel binnen de case-organisatie (om de praktische relevantie te borgen), als binnen de Open Universiteit doormiddel van de presentatie van de thesis (voor het delen van de wetenschappelijke bijdrage). Als het afstudeerverslag voldoende wordt beoordeeld, wordt deze door de Open Universiteit gepubliceerd.

3.3. Reflectie t.a.v. Rigor en Relevance

Design Science Research (DSR) creëert en evalueert IT-artefacten die zijn bedoeld om geïdentificeerde organisatorische problemen op te lossen. Binnen DSR worden Rigor (het wetenschappelijke domein) en Relevance (het organisatie domein) met elkaar verbonden (Hevner et al., 2004). Rigor wordt bereikt door bestaande wetenschappelijke foundations en methodieken correct toe te passen. In de DSR worden deze methoden voornamelijk gebruikt om de kwaliteit en effectiviteit van artefacten te evalueren, daarnaast kunnen ook empirische technieken worden gebruikt (Hevner et al., 2004). Relevance wordt bereikt het door creëren van een toepasbaar artefact welke binnen een zakelijke omgeving getoetst kan worden (Hevner et al., 2004). In onderstaand figuur 11 is schematisch het verband weergegeven tussen het DSR, enerzijds het wetenschappelijke domein verbonden door de Rigor cycle en anderzijds het organisatie domein

verbonden door de Relevance cycle (Hevner, 2007).



Figuur 11 Design Science Framework (Hevner et.al. 2004)

Om de Rigor cyclus van het artefact te waarborgen, maakt deze DSR gebruik van een theoretisch kader gebaseerd op peer reviewed wetenschappelijke literatuur. Daarnaast zijn binnen de case organisatie meerdere bronnen geraadpleegd (zowel door middel van (groeps)interviews als deskresearch) om een chain of evidence te creëren.

Om de relevance cyclus van het artefact te waarborgen, zijn de te interviewen stakeholders geselecteerd op basis van hun professionaliteit omtrent data science projecten en zijn zij vooraf gebriefd over het onderzoek om te zorgen dat zij goed geïnformeerd zijn en zich alvast kunnen voorbereiden. Vervolgens is ook tijdens groepsinterview sessies de toepasbaarheid en de effectiviteit van het artefact voor data science projecten onderzocht en bevestigd.

Het DSR-onderzoek is volledig en effectief wanneer het voldoet aan de vereisten en beperkingen van het probleem dat het moest oplossen. De resultaten van de evaluatie van het artefact zijn opgenomen in hoofdstuk 4

4. Resultaten

In dit hoofdstuk wordt het ontwikkelende artefact gepresenteerd. Vervolgens wordt het model gedemonstreerd en worden de gegevens geëxtraheerd. De laatste stap is het evalueren van de gegevens en het presenteren van de resultaten.

4.1. Design

In het theoretisch framework wordt uitgewerkt hoe de verschillende projectmethoden/processtappen bijdragen/invulling geven aan de belangrijkste projectkenmerken Project context, Analytics context, Data context, Team context en Organizational context welke zijn uitgewerkt in hoofdstuk 2.3.2.

Project context

Binnen de project context worden twee typen project onderscheiden, routinematige en éénmalige projecten. Bij de routinematige projecten worden meestal modellen of rapporten gemaakt om een bedrijfsproces te ondersteunen, (Ahangama & Poo, 2015; Li et al., 2016). Als het model hergebruikt dient te worden voor volgende projecten is een automate stap nodig (Mariscal et al., 2010) . Service georiënteerde data analyses zijn vaak onderdeel van de bedrijfsprocessen. Als deze informatie ook nog real-time beschikbaar moet zijn (criticality of timelines volgens Saltz et al 2017) is een support of maintenance stap nodig. In principe zijn deze stappen zowel in een watervalmethode, als in Agile methoden als Scrum of Kanban toe te voegen. Bij de eenmalige projecten worden meestal ad-hoc vragen of problemen onderzocht. Deze projecten zijn vaak minder gestructureerd dan routinematige projecten (Saltz & Shamshurin, 2015). Deze projecten zijn lastiger planbaar en het is ook lastiger vooraf een sluitende business case te definiëren. Tijdens deze projecten zullen regelmatig iteraties plaats vinden, waardoor een watervalmethode minder geschikt is. Eén van de Agile methoden zou wel geschikt zijn omdat hier continue op eventuele veranderingen in kan worden gespeeld. Kanban lijkt dan meer geschikt aangezien het voor een sprintlengte bij Scrum wenselijk is dat werk planbaar is (Saltz et al., 2018).

Data Context

Big Data projecten bevatten grote hoeveelheden (Volume) gevarieerde (Variety) data welke in korte tijd (Velocity) wordt geproduceerd. Deze informatie dient tijdens de data preparation fase verwerkt te worden. Tijdens deze projecten is het belangrijk snel te kunnen schakelen bij veranderingen, hiervoor is een agile projectaanpak gewenst (Gao et al., 2015; Saltz, 2015; Saltz & Shamshurin, 2016). Daarnaast zijn er iteratieve mogelijkheden nodig om deze Big Data op de juiste manier te interpreteren. Kanttekening is, als één van de factoren te groot is om in één sprint te passen, Scrum minder geschikt is. Bij projecten met meer gestructureerde data kan wel een gestandaardiseerd proces worden gevolgd, maar liefst wel met iteratieve stappen om in te kunnen spelen op eventuele veranderingen en toepassen/verwerken van evaluaties. Zowel een hybride projectvorm (Waterval + iteraties of Agile) als agile methoden als Scrum en Kanban kunnen hiervoor gebruikt worden (Gao et al., 2015; Saltz, 2015; Saltz & Shamshurin, 2016).

Analytical context

Binnen de analytics context worden twee typen analyses onderscheiden, hypothese genererende en hypothese testende analyses. Bij een open analyses (genereren van hypotheses), zijn iteratieve stappen nodig om het proces richting te geven, hiervoor is een waterval gedreven model minder geschikt. Ook kennen deze projecten meer uitdagingen met betrekking tot tijdsplanning dan met gerichte analyses (testen van hypotheses) (Saltz, 2017), hoe meer de projectmodellen geschikt zijn voor open analyses, hoe lastiger ze vooraf planbaar zijn. Agile methoden als Scrum en Kanban

bieden de mogelijkheid voor iteraties wel en zijn daarom meer geschikt. Een hybride vorm met voldoende iteratiemogelijkheden is ook mogelijk. Bij gerichte analyse (testen van hypotheses) is vooraf bekend welke informatie men wil. Dit maakt het project makkelijker planbaar. Hier volstaat zowel een waterval gedreven model als één van de Agile methoden. Als een projectmodel herbruikbaar moet zijn is een Automate stap nodig, ook is dan de Support / Maintenance stap gewenst (Mariscal et al. 2010)

Team context

Een specifiek projectmodel geeft geen invulling aan het projectkenmerk Manager Experience, andersom zal de achtergrond en ervaring van de manager wel mede bepalen voor welke projectmethode hij kiest. Ook voor Virtuality of Team telt dat geen specifiek projectmodel invulling geeft aan dit projectkenmerk, andersom zal de achtergrond en ervaring van team wel mede bepalen voor welke projectmethode men kiest. Mariscal (2010) heeft wel een aparte stap voor Human resource identification (bij de anderen valt deze onder business of domain understanding), maar daar wordt voornamelijk gesproken hoe een team samen te stellen Gao et al., (2015) geven aan dat een multidisciplinair team gewenst is. Andersom kan er bij het samenstellen van een team wel rekening worden gehouden met de ervaring met een bepaalde projectmethode. Samengevat zijn deze teamkenmerken niet bepalend voor de projectkeuze, maar zal als er ervaring is met slechts 1 projectmethode dit de keuze wel beïnvloeden. Bij agile projecten met multidisciplinaire teams is communicatie over taken en inzichtelijkheid van voortgang belangrijk, hierbij past ook een agile projectmethode (Batra, 2018).

Organizational Context

Organization Culture & Process – Als de focus van een organisatie op processen ligt, zullen die meegenomen worden in de keuze voor het procesmodel. Een organisatie kan uiteindelijk voorkeur hebben voor elke procesvorm, zowel watervalmethoden als Agile methoden als Scrum of Kanban. Als een vooraf vastgesteld analyse proces gewenst is, zijn Scrum en Kanban eventueel minder geschikt. Organization Culture & ROI – Als een organisatie een duidelijke vooraf gedefinieerde impact (bijvoorbeeld door middel van een business case) verwacht, moet deze stap opgenomen zijn in het procesmodel, alle procesmodellen bevatten deze mogelijkheid wel in de Business understanding fase, ook bij Scrum als Kanban is deze stap toe te voegen. Wel is het zo dat open analyses (zoals bijvoorbeeld door Scrum en Kanban) lastig planbaar zijn en daarvan dus lastig is vooraf een business case te maken. Aangezien de methoden Scrum en Kanban als eigenschap hebben lastig planbaar te zijn, zijn die minder geschikt als de organisatie een vooraf gedefinieerde impact wenst.

4.1.1. Raamwerk

Onderstaand framework geeft een schematisch overzicht van de eerder benoemde verbanden tussen de projectkarakteristieken en de projectmethoden/processtappen

			CRISP-DM	Aanvullende Thema's				Aangepast Proces		
				Life Cycle Selection	Conceptualization	Automate	Support / Maintenance	Hybride (Waterval + iteraties of Agile)	Agile Scrum	Agile Kanban
Project scope	Type project	Routinematig				x	x	x	x	x
		Eenmalig	x					x	x	x
Analytics context	Type of Analysis	Hypothese generating		x			x	x		x
		Hypothese testing	x			x	x	x	x	x
Data Context	Big Data							x		x
	Gestructureerd		x			x	x	x	x	x
Team context	Manager Experience		x					x	x	x
	Team Experience		x					x	x	x
Organizational Context	Organization Culture & Process		x				x	x	x	x
	Organization Culture & ROI		x				x			

Table 1 Framework met verbanden tussen projectkarakteristieken en projectmethoden/processtappen

4.2. Demonstration

Het doel van de demonstratie is om conform de Design Science methode, het raamwerk te toetsen in de praktijk. Dit is gedaan door het uitvoeren van een case study bij data science specialisten binnen een case organisatie.

4.2.1. Case organisatie

De organisatie waar het onderzoek wordt uitgevoerd is een midden-grote fabrikant van 3D printers. Deze organisatie is opgericht in 2011 en heeft ca. 400 mensen in dienst. Naast het ontwikkelen, produceren en verkopen van 3D printers, is deze organisatie ook de leverancier van een open source applicatie voor het printen van 3D objecten. Deze applicatie heeft momenteel ruim 3 miljoen hits per maand (april 2020).

De organisatie heeft sinds kort (zomer 2020) een nieuw Data & Metrics team, daarnaast binnen marketing een afdeling Research & Analytics (voornamelijk gericht op markt analyse), deze afdeling maakt ook gebruik van Big Data. Overige interne data analyses worden door de operationele afdelingen zelf gedaan. De reden van het ontstaan van het Data & Metrics team is dat er geen eenduidige bron van informatie was welke door de afdelingen kon worden gebruikt voor het bouwen van hun rapporten en dat er een versplintering van data analyse rapporten binnen de organisatie was.

4.2.2. Het onderzoek

Het onderzoek bestaat uit meerdere stappen. Ronde één met individuele semi-structured interviews, gevolgd door een documentonderzoek en een tweede ronde met een focusgroep interview . Daarna ronde drie met een tweede focusgroep interview .

Ronde één

De eerste ronde van het onderzoek is uitgevoerd door het afnemen van oriënterende semi-structured interviews bij vijf experts binnen alle drie bovengenoemde groepen. Deze personen zijn gekozen na overleg met de Chief Information Officer op basis van hun ervaring met data analyse projecten:

- Product Owner Data & Metrics
- Director Global Research & Analytics
- Content Marketing Specialist Marketing Research & Analytics
- Business Manager Customer Service (als analist van de afdeling Service)
- Project Engineer Research & Development (als analist van de afdeling R&D)

De interviewvragen zijn gebaseerd op het theoretisch framework uit hoofdstuk 2.3.3. Doel van de eerste ronde is om zicht te krijgen in de ervaring van de geïnterviewden en de organisatie over het gebruik van projectmethoden en procestappen voor data-science projecten. Daarnaast ook om inzicht te krijgen in de karakteristieken van de diverse data science projecten binnen de organisatie. Na de eerste ronde wordt de verkregen informatie verwerkt en getoetst op het ontwikkelde framework uit hoofdstuk 4.1.6.

Documentonderzoek

Na de eerste ronde heeft ook een documentonderzoek plaatsgevonden. De documentatie bestond uit presentaties over de IT software delivery organization, introductie van het nieuwe Data & Metrics team, de toekomstige Data Architecture, plus operational procedures van de afdelingen rondom data analyse projecten (voor zover beschikbaar). De inhoud van de documentatie heeft bijgedragen aan en input gegeven voor de evaluatie van het framework. Aanvullende reden voor het documentonderzoek was een chain of evidence te creëren door het raadplegen van meerdere bronnen.

Ronde twee

Na de eerste ronde interviews en het documentonderzoek, was er nog onvoldoende informatie beschikbaar als input voor een goede formatieve evaluatie van het framework, daarom is een tweede ronde gehouden. Deze tweede ronde bestaat uit een focusgroep interview met als doel dieper in te gaan op het verband tussen de processtappen en de projectkarakteristieken. Dit focusgroep interview is gehouden met 3 experts:

- Project Engineer Research & Development (als analist van de afdeling R&D)
- Data Engineer Data & Metrics
- Global Business Analist Sales & Marketing

Het voordeel van een focusgroep interview is dat de geïnterviewden samen kunnen discussiëren over de vragen en antwoorden. Hierdoor kunnen ze ook elkaar bevragen en daarmee gezamenlijk tot een volledig(er) antwoord komen (Saunders, et al., 2016). De interviewvragen/ -onderwerpen zijn gebaseerd op de informatie uit de eerste ronde en op het framework uit hoofdstuk 2.3.3. Om er zeker van te zijn dat er geen onduidelijkheden waren over de projectmethoden, processtappen en de projectkarakteristieken, is deze ronde gestart met een korte presentatie waarin de gebruikte termen en de structuur van het framework uit hoofdstuk 2.3.3 is uitgelegd. Tijdens de focusgroep discussie zijn de vijf groepen projectkarakteristieken doorgelopen en de verbanden met de projectmethode/ processtappen bediscussieerd. Het doel van deze ronde was meer diepgang en overeenstemming te krijgen over de gewenste processtappen. Kijkend naar de

projectkarakteristieken van de organisatie, wat zouden dan logische processtappen zijn? De input uit deze tweede ronde is, samen met de eerste ronde en het documentonderzoek, gebruikt voor de formatieve evaluatie en het verbeterde framework in hoofdstuk 4.3.2.

Ronde drie

De derde ronde bestaat uit een focusgroep interview met als doel het verbeterde framework uit hoofdstuk 4.3.2 te bespreken en te valideren. Dit focusgroep interview is gehouden met 3 specialisten:

- Project Engineer - Research & Development (als analist van de afdeling R&D)
- Data Engineer - Data & Metrics
- Global Busines Analist - Sales & Marketing

De interviewvragen/ -onderwerpen zijn gebaseerd op de informatie uit de eerste en tweede ronde en op het framework. Tevens is er gebruik gemaakt van een discussiemodel (zie bijlage 2) met processtappen als basis voor de verificatie. Tijdens deze sessie zijn de extra proces stappen t.o.v. CRISP-DM besproken (en hun toegevoegde waarde), plus welke processtappen bij welke project karakteristieken passen. Doel van de derde ronde is om de toepassing op het verbeterde framework in hoofdstuk 4.3.2 te toetsen met de geïnterviewden en als basis voor de summatieve evaluatie.

Een uitgebreider overzicht van de interviews is terug te vinden in bijlage 2 Interview Guide.

Door de Corona crisis was het niet mogelijk de interviews face to face te organiseren. Daarom is gekozen deze via Microsoft Teams te houden. Microsoft Teams heeft ook de functionaliteit de interviews op te nemen om te gebruiken voor de transcripties. Van elk interview zijn transcripties gemaakt welke worden gebruikt voor de evaluatie.

4.2.3. Verwerking

Nadat de interviews getranscribeerd zijn, zijn ze gecodeerd met behulp van NVivo volgens het codeerschema in bijlage 3. De transcripties zijn geanonimiseerd en gevalideerd door de deelnemers. Het codeerschema is opgesteld op basis van de projectkarakteristieken en de projectmethoden van het raamwerk. De codes zijn vervolgens gegroepeerd en gebruikt om verbanden zichtbaar te maken tussen de projectkarakteristieken en kenmerken van een procesmodel. Dit vormt een overzicht welke gebruikt is als basis voor de evaluatie.

4.3. Evaluation

De evaluatie bestaat uit twee stappen. Allereerst de Formatieve evaluatie (tijdens de eerste ronde, oriënterende interviews, het documentonderzoek en de tweede ronde, het eerste verdiepend focusgroep interview). De formatieve evaluatie heeft als doel het framework uit hoofdstuk 4.1.6 te evalueren en eventueel te verbeteren. Vervolgens de Summatieve evaluatie (tijdens de derde ronde, in een tweede verdiepend focusgroep interview) waarin het verbeterde framework uit hoofdstuk 4.3.2 wordt besproken, getoetst en geëvalueerd.

4.3.1. Formatieve evaluatie

De formatieve evaluatie bestaat uit drie stappen. Allereerst een eerste ronde, oriënterende interviews met experts, vervolgens het documentonderzoek en als derde een tweede ronde, het eerste verdiepend focusgroep interview.

Algemeen

Een vaste bedrijfsbrede leidraad voor data science projecten ontbreekt. De meeste teams werken niet volgens een bepaalde projectmethode, alleen het Data & Analytics team gebruikt de Kanban

methode voor zijn planning. Het nieuwe Data & Metrics team gaat ook werken in sprints om zo per sprint af te kunnen stemmen wat er geleverd gaat worden. Sommige teams hanteren een ad-hoc aanpak, andere teams hebben wel een leidraad, maar deze is niet bedrijfsbreed geïntroduceerd. De stappen van CRISP-DM worden vaak wel gebruikt, maar vaak onbewust en vaak ook als input voor Agile Scrum of Kanban (of een hybride versie). Het belangrijkste voordeel wat wordt gezien is dat hierdoor stappen gecombineerd kunnen worden, of parallel uitgevoerd, waardoor er niet/minder op elkaar gewacht hoeft te worden (bij een waterval methode volgen stappen elkaar op). Life Cycle selection wordt vooralsnog als onbekend en niet waardevol geacht.

Project context

De meeste projecten binnen de case organisatie zijn routinematige projecten. Ondanks de routinematigheid werken de meeste teams niet volgens één bepaalde projectmethode, zowel (stappen van) CRISP-DM, Scrum, Kanban, als Hybride vormen worden gebruikt. Alle teams geven aan dat de Business understanding vs problem specification fase de lastigste, maar eigenlijk wel de belangrijkste fase is, bij zowel routinematige als eenmalige projecten. Bij met name het Data & Analytics team wordt daarom ook extra aandacht aan deze fase gegeven. Andere teams kiezen er vaak voor meerdere iteraties en evaluaties op het model te doen, om zo de vraag en het antwoord verder te specificeren. Maintenance wordt bij routinematige projecten gezien als een belangrijke stap, maar is nog onvoldoende belegd. In de praktijk wordt deze vaak uitgevoerd door de business owner. Automate wordt nog weinig toegepast, de meeste analyses worden vanaf scratch opgebouwd. Voor 2021H2 wordt gebouwd aan een datalake welke voor zowel routinematige als eenmalige projecten kan worden gebruikt. Mede hiervoor is een Data & Metrics team opgericht.

Verbanden voor framework:

- Routinematige projecten
    - Projectmethoden: CRISP-DM, Scrum, Kanban, Hybride
    - Processtappen: Business understanding vs problem specification, Automating, Maintanance
- Eenmalige projecten
    - Projectmethoden: Scrum, Kanban, Hybride,
    - Processtappen: Business understanding vs problem specification, Automating

Data context

Binnen de case organisatie wordt voornamelijk gewerkt met interne data afkomstig uit de verschillende software pakketten (gestructureerde data). Alleen het Marketing Research & Analytics team werkt ook met externe data (en Big data zoals Social media data), voornamelijk voor hun marktonderzoek. De stappen Data understanding en data preperation kosten soms zoveel tijd dat deze niet in 1 sprint kunnen worden uitgevoerd. In dat geval wordt Agile Scrum als een minder wenselijke methode gezien. Momenteel wordt voor elk project een nieuwe dataset bepaald en bij elkaar gezocht, bestaande datasets worden niet opnieuw gebruikt, er is dus nog geen sprake van Automating. De case organisatie erkend wel het belang van Automating en wil dit zoveel mogelijk laten borgen door het nieuwe Data & Metrics team. Wat binnen de case organisatie opvalt is dat er rond de stappen Data understanding en Data preperation vaak een aanvullende actie plaats vindt, namelijk het genereren van de juiste Data(set). De benodigde data is niet altijd beschikbaar, aan het project wordt het onderzoek naar de juiste data dan toegevoegd. Recent is het Data & Metrics team opgericht, zij moeten er voor zorgen dat zoveel mogelijk data gecentraliseerd wordt in een data lake en beschikbaar gesteld moet worden, met name om te zorgen voor one single source of truth, dit ter bevordering van de data understanding en data preparation stap. Dit betreft voornamelijk de gestructureerde data. Inmiddels is de wens voor conceptualizatie ontstaan. Waar deze stap

oorspronkelijk was bedoeld als het toepassen van een conceptueel datamodel voor de gezondheidszorg (in verband met privacy), wil de case organisatie deze gaan toepassen op projecten voor met name de Marketing afdeling waar gewerkt wordt met (Big data) Social media data en persoonsgegevens. Maintenance is belangrijk voor zowel gestructureerde als Big data.

Verbanden voor framework:

- Big Data
    - Projectmethoden: Kanban, Hybride
    - Processtappen: Business understanding vs problem specification, Automating, Maintenance
- Gestructureerde data
    - Projectmethoden: CRISP-DM, Hybride
    - Processtappen: Business understanding vs problem specification, Automating, Maintenance

Analytics context

Binnen de case organisatie worden vooral hypothese testing analyses gedaan, alleen de afdeling Data & Analytics doet ook regelmatig hypothese generating analyses. Hypothese testing analyses worden vaak als routinematig of gestandaardiseerd gezien, desondanks hebben de meeste afdelingen nog geen vaste projectmethode voor deze routinematige analyses. Sommigen gebruiken de stappen van CRISP-DM (eventueel onbewust), anderen gebruiken Agile vormen als Scrum of Kanban. Bij korte doorlooptijd wordt Scrum minder geschikt geacht. Voor de hypothese generating wordt meer vastgehouden aan een vaste projectmethode, namelijk in de vorm van CRIP-DM, inclusief de iteratieve stappen. Dit komt omdat deze voornamelijk worden uitgevoerd door het Data & Analytics team, die wel een vaste projectmethode hanteren, zij gebruiken weer geen Scrum en Kanban (maar geven aan dat dit wel zou kunnen). Sowieso vindt iedereen iteratieve mogelijkheden noodzakelijk, voor zowel hypothese testing, als generating projecten. Problem definition wordt binnen alle afdelingen, voor zowel hypothese testing als generating projecten, als één van de belangrijkste stappen voor de data analyse projecten gezien, business understanding is daarbij erg belangrijk. Voor hypothese testing analyses wordt vaak dezelfde (interne) data gebruikt en als het resultaat van de analyse (het model of rapport) succesvol is, worden vaak periodieke rapportages gewenst. Ondanks het routinematige gebruik van dezelfde data, wordt het data preperation deel elke keer opnieuw uitgevoerd. Met het toekomstige data lake zal deze stap in ieder geval vereenvoudigen. Voor hypothese generating wordt regelmatig gebruik gemakt van Big Data, zowel een traditioneel watervalmodel als Scrum worden minder geschikt geacht (zie ook data context) Ook is er, ondanks de routinematigheid van de hypothese testing analyses, nog onvoldoende nagedacht over het automatiseren en het beleggen van onderhoud op de ontwikkelde modellen en rapporten. Het belang wordt wel onderkent, de operationele belegging ontbreekt. Criticality of Timelines als in realtime beschikbaarheid is binnen de huidige projecten niet relevant, het nieuw te bouwen data lake zal eventuele vertraging die nu daarin ervaren wordt gaan verhelpen.

Verbanden voor framework

- Hypothese testing:
    - Projectmethoden: CRISP-DM met iteratieve mogelijkheden, (Scrum), Kanban
    - Processtappen: Busines understanding Problem definition, Data preperation Automating, Maintenance
- Hypothese generating
    - Projectmethoden: CRISP-DM, (Scrum, Kanban), Hybride

- Processtappen: Busines understanding Problem definition, Data preperation

Team context

De case organisatie kent geen echte data science teams. Wel heeft de afdeling marketing een eigen Marketing Research & Analytics team, zij hebben ook de meeste ervaring met data science projecten en we zien dat daar het best gebruik wordt gemaakt van standaard processtappen, zij zijn bekend met CRISP-DM en passen dit ook gedeeltelijk toe (wel met iteratieve stappen). De overige data science projecten worden meestal uitgevoerd door kleine operationele teams, dit zijn meestal multidisciplinaire teams waar 1 of meerdere personen expertise/gevoel met data science hebben. Zij hebben over het algemeen weinig/geen ervaring met /kennis van de mogelijke processtappen van data science en gebruiken die ook niet bewust. Zij vallen terug op de kennis van algemeen projectmanagement, er wordt bijvoorbeeld wel gebruik gemaakt van Agile scrum en Kanban, maar zonder de ervaring dit toe te passen op data science projecten. Wel wordt Agile scrum en Kanban als voordeel gezien in verband met korte en snelle communicatie.

Verbanden voor framework

- Manager Experience
  - Projectmethoden: CRISP-DM, Scrum, Kanban
  - Processtappen: Geen specifieke processtappen
- Team Experience
  - Projectmethoden: CRISP-DM, Scrum, Kanban
  - Processtappen: Geen specifieke processtappen
- Size of team
  - Projectmethoden: Scrum, Kanban
  - Processtappen: Geen specifieke processtappen

Organizational context

De case organisatie is op moment van het onderzoek nog weinig bezig met de ROI van de data science projecten. Dit wordt mede veroorzaakt doordat de meeste data science projecten binnen de operationele teams vallen en niet zijn gecentraliseerd. Over het algemeen zijn de projecten ook vrij klein van omvang en kortlopend. De case organisatie is enigszins proces gedreven. Binnen een groot aantal operationele afdelingen wordt gewerkt volgens de Scrum methodiek (d.m.v. sprints), een aantal andere afdelingen werkt met Kanban borden voor onder andere overzicht en structuur. Als specifieke data science processtappen worden gebruikt, worden die vaak in één van deze methoden toegepast. Er is vooralsnog geen algemene procesmethode die door de hele organisatie wordt toegepast. Dit is ook zichtbaar binnen de diverse data science projecten, een vaste aanpak ontbreekt. De organisatie heeft geen algemene Data science teams, de analisten van de diverse afdelingen werken in de operatie. Wel is er een algemeen Data & Metrics team welke organisatiebreed opereert. Tevens is er een overlegstructuur met de analisten van de operationele afdelingen en het Data & Metrics team.

Verbanden voor framework

- Organizational Culture & Process,
  - Projectmethoden: Scrum, Kanban
  - Processtappen: Support / Maintenance is geen onderdeel van de standaard processen
- Organizational Culture & ROI → Organisatie is niet ROI gedreven

4.3.2. Geupdate Framework

Na de formatieve evaluatie zijn er een aantal punten uit het oorspronkelijke framework aangepast. De aanpassingen zijn eerst in onderstaand overzicht beschreven en daarna verwerkt in het aangepaste onderstaande framework.

Thema	Aanpassing
Algemeen	Life Cycle selection wordt als onbekend en niet waardevol geacht
Project context	Bij zowel routinematige als éénmalige projecten is een extra stap voor Problem definition gewenst
Data context	Bij zowel Hypothese testing als generating projecten is een extra stap voor Problem definition gewenst
Data context	Bij projecten met Big Data wordt Agile Scrum minder geschikt geacht.
Data context	Bij (Big data) projecten met Social media en persoonsgebonden data is conceptualization gewenst
Team context	Bij kleine team is de voorkeur voor Agile Scrum of Kanban
Organization Culture & ROI	Hier wordt binnen de case organisatie geen rekening gehouden, dus kon niet worden beoordeeld

Table 2 Aanpassingen framework

Nieuwe framework

Toevoegingen in groen, verwijderd in rood

			CRISP-DM	Aanvullende Thema's				Aangepast Proces		
				Problem definition	Life Cycle Selection	Conceptualization	Automate	Support / Maintenance	Hybride (Waterval + iteraties of Agile)	Agile Scrum	Agile Kanban
Project scope (1)	Type project	Routinematig		x			x	x	x	x	x
		Eenmalig	x	x					x	x	x
Analytics context (2)	Type of Analysis	Hypothese generating		x	x				x		x
		Hypothese testing		x			x	x	x	x	x
Data Context (3)	Big Data					(x)	x		x		x
	Gestructureerd		x				x	x	x	x	x
Team context (4)	Manager experience		x						x	x	x
	Team experience		x						x	x	x
	Size of team									x	x
Organizational Context (5)	Organization Culture & Process	Process Oriented	x					x	x	x	x
	Organization Culture & ROI		x			x		x			

Table 3 Aangepaste framework

4.3.3. Summatieve evaluatie

De summatieve evaluatie heeft plaatsgevonden in het tweede focusgroep interview met de specialisten. Deze sessie had als doel om het ontworpen framework inhoudelijk te bespreken en tevens te evalueren op praktische bruikbaarheid en volledigheid. Het framework is tijdens deze studie nog niet toegepast in de praktijk, waardoor het nog niet getoetst kan worden op effectiviteit. De praktijktoets is dan ook een aanbeveling voor vervolgonderzoek. Uiteindelijk is het raamwerk daardoor getoetst op vier van de vijf criteria van Prat et al. (2014) :

Goal

Tijdens de tweede focusgroep sessie is de relevantie en bruikbaarheid van het aangepaste framework besproken en bevestigd. Hierbij is niet het framework zelf getoond, maar zijn alle individuele thema's van de projectkarakteristieken en het discussiemodel uit bijlage 3 besproken. Dit discussiemodel was samengesteld op basis van de stappen uit CRISP-DM, aangevuld met de relevante processtappen uit de expert interviews en de eerste focusgroep sessie. De relatie tussen de projectkarakteristieken en de processtappen uit het discussiemodel werden tijdens deze sessie bevestigd. Tevens werden de individuele thema's van de projectkarakteristieken bevestigd. Na de sessie werd aangegeven dat het discussiemodel, inclusief de optionele stappen, bruikbaar is in de praktijk.

Environment

Tijdens de tweede focusgroep is het discussiemodel getoond en besproken, inclusief de individuele thema's van de projectkarakteristieken welke tijdens de expert interviews en de eerste focusgroep sessie uitgelicht waren. Tijdens de eerste focusgroep sessie is een korte presentatie gegeven om de individuele projectkarakteristieken en processtappen uit te leggen. Onduidelijkheden zijn direct uitgelegd. Bovengenoemde projectkarakteristieken en processtappen zijn getoetst op consistentie en door de deelnemers bevestigd tijdens de sessie.

Structure

Tijdens beide focusgroep sessies zijn de individuele projectkarakteristieken en processtappen besproken en getoetst bij de deelnemers. Hierbij is gebruik gemaakt van het framework op uit hoofdstuk 2.3.3. Onduidelijkheden zijn toegelicht. Tevens is besproken of er nog onderdelen misten, hier werd aangegeven dat het overzicht in hun optiek compleet is.

Activity

Niet getoetst tijdens dit onderzoek

Evolution

Tijdens beide focusgroep sessies is besproken of het discussiemodel en de thema's toepasbaar zijn op de verschillende soorten projecten (of projecten met verschillende eigenschappen), dit werd ook bevestigd. De combinatie van de processtappen en de projectkarakteristieken zijn na de eerste focusgroep sessie verwerkt in het nieuwe framework in hoofdstuk 4.3.2. Door de bevestiging dat voorgenoemde combinatie van de procestappen en de projectkarakteristieken compleet was en bruikbaar voor verschillende soorten projecten (of projecten met verschillende eigenschappen), kunnen we concluderen dat het framework voldoet aan de robuustheidstoets.

5. Conclusies en aanbevelingen

In dit hoofdstuk worden de conclusies van het onderzoek en de bijdrage aan de wetenschap besproken, plus aanbevelingen voor de praktijk en vervolgonderzoek gedaan.

5.1. Conclusies

In de literatuurstudie worden de deelvragen over de processtappen en de projectkarakteristieken beantwoord, deze vormen tezamen het theoretisch framework. Het Design Science research had als doel een framework te creëren welke door organisaties gebruikt kon worden om de juiste processtappen/-methoden te selecteren, passende bij de karakteristieken van hun data science projecten. In het framework is inzichtelijk gemaakt welke processtappen en thema's belangrijk zijn voor en succesvolle projectaanpak, afgestemd op het type data science project, of specifieke projectkarakteristieken. Hierdoor wordt de onderzoeksvraag van dit rapport beantwoord:

Hoe kunnen projectmethoden/processtappen worden gebruikt om een data science project succesvol te maken?

Kijkend naar de uitkomst van het onderzoek kunnen we de volgende conclusies trekken en toepassen op het framework (Hoofdstuk 4.3.2). De case organisatie heeft geen vaste manier van projectaanpak voor data science projecten. De processtappen van CRISP-DM worden grotendeels wel gevolgd, maar niet altijd bewust of consequent. Deze stappen worden eventueel gecombineerd met Agile methoden als Scrum of Kanban. De keuze van projectmethode hangt voornamelijk af van de (1) Project context en de (2) Analytics context. Bij zowel routinematige als éénmalige projecten en Hypothese testing als generating projecten is binnen Business understanding de extra stap Problem definition gewenst. Ook is er binnen beide typen project behoefte aan iteratieve stappen. Hierdoor is een strikt watervalgedreven proces niet geschikt, wel kan CRISP-DM, met oorspronkelijke iteraties, Scrum, Kanban of een hybride vorm worden gebruikt. Uitzondering zijn projecten met Big data waar Agile Scrum door de afbakening van de sprints niet geschikt wordt geacht. De meeste projecten zijn hypothese testende routinematige projecten die vaak dashboards als model voor de business owner opleveren. Voor deze projecten is de stap Maintenance voor het onderhoud een must. Deze projecten kunnen van binnen de (3) Data context zowel gestructureerde als Big data gebruik maken. De stap Automating is bij deze projecten ook een wens, enerzijds voor de snelheid en handelbaarheid van oplevering van de data en het model, anderzijds als basis voor doorontwikkeling voor het model. Tot slot is de wens Conceptualization toe te passen voor sommige Big data projecten (3) die gebruikmaken van Social media data en persoonsgegevens. De keuze voor een projectmodel wordt nu nog vaak gemaakt op basis van de ervaring van de manager of het team (4), hier kan zowel CRISP-DM, Scrum of Kanban of een hybride versie worden gekozen. Binnen de case organisatie zijn geen specifieke data science teams, maar worden de meeste data science projecten uitgevoerd door kleine multidisciplinaire teams (4), voor wie flexibiliteit en open communicatie belangrijk zijn. Door Agile methoden kunnen zij werkzaamheden beter combineren of parallel uitvoeren, wat de doorlooptijd van een project positief beïnvloed. Binnen de case organisatie is er wel focus op procesgericht (5)werken, maar is er geen wens voor een specifiek proces, daardoor kan zowel CRISP-DM, Scrum of Kanban worden gebruikt. Wel wordt Scrum al door meerdere teams gebruikt, waardoor de meeste medewerkers bekend zijn met deze methode. De case organisatie kent geen focus op ROI (5), waardoor dit geen invloed heeft op de keuze voor processtappen.

5.2. Bijdrage aan de wetenschap

Dit rapport voegt twee punten toe aan de wetenschap van data science. Dit is mede te danken aan de structuur van het Design Science research. Allereerst geeft het rapport inzicht en overzicht van de verschillende projectkarakteristieken van data science projecten en verschillende projectstappen die hiervoor gebruikt kunnen worden. Dit inzicht is verkregen door een wetenschappelijk literatuuronderzoek. Hier is aangegeven welke thema's uit recentere onderzoeken toegevoegd dienen te worden aan de processtappen CRISP-DM om een compleet overzicht van proces methoden te krijgen. Ten tweede heeft het onderzoek geresulteerd in een framework, waarin op basis van de karakteristieken van een data science project, gekozen kan worden voor de juiste processtappen om het data science project op de juiste manier in te richten en daarmee de slagingskans te vergroten. Dit framework is gebaseerd op literatuuronderzoek, aangevuld door een case study binnen een case organisatie. Tijdens deze case study is het framework ook getoetst op bruikbaarheid en volledigheid. Dit framework kan als basis gebruikt worden voor vervolgonderzoek.

5.3. Aanbevelingen voor de praktijk

Het tijdens het focusgroep interview besproken framework is bruikbaar in de praktijk. Dit is tijdens deze sessie besproken en bevestigd. Tijdens dit onderzoek is het framework niet meer in de praktijk getest op effectiviteit en prestatie, maar naar de mening van de deelnemers wel toepasbaar. Om de effectiviteit alsnog te bevestigen is de aanbeveling het framework te introduceren aan de overige analisten binnen de organisatie en hier een pilot periode gebruik van te maken. Tijdens deze praktijktoets is het belangrijk op basis van de projecteigenschappen van de data analyse projecten de processtappen uit het framework te selecteren. De focus zal, zeker in het begin, op de eerste processtappen moeten liggen, Problem formulating, Business understanding en Data understanding (dit werd als zwaartepunt door de deelnemers aangegeven). Regelmatige evaluatie en iteratie tijdens deze stappen wordt ook zeer aanbevolen. Op dit moment is er uberhaupt nog geen éénduidige projectaanpak voor data analyse projecten, het is aan te bevelen zo snel mogelijk met de praktijktoets te starten om ook éénduidigheid te krijgen.

5.4. Limitatie en aanbevelingen voor verder onderzoek

Limitaties:

- Door de gerichte case study binnen een case organisatie is generalisatie niet altijd goed mogelijk.
- De case organisatie is nog redelijk onvolwassen op het gebied van data science projecten. Andere organisaties die al verder in hun ontwikkeling zijn hebben mogelijk andere wensen of eisen aan het framework.
- Gedurende de studie hadden 2 specialisten van de afdeling Marketing Research & Analytics de case organisatie verlaten, waardoor veel kennis verloren is gegaan.
- Het ontworpen framework is nog niet getest in de praktijk. Bij praktisch gebruik kan de kwaliteit en bruikbaarheid nogmaals bevestigd worden. (of er kan worden geconcludeerd dat er nog een iteratie nodig is)

Aanbevelingen

- Voor betere generalisatie is een vergelijkbaar onderzoek (en dan met name de demonstratie en de evaluatie) binnen andere organisaties aan te bevelen
- Het is aan te bevelen het framework in de praktijk te toetsen om daarmee kwaliteit en bruikbaarheid te toetsen.

- Het is aan te bevelen het framework ook te toetsen binnen een organisatie welke verder in hun data science ontwikkeling zijn, omdat zij mogelijk extra eisen stellen aan het framework.

Referenties

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

Anderson, C. (2015). Chapter 1. What Do We Mean by Data-Driven? In *Creating a data-driven organization* (p. 285). Retrieved from https://www.oreilly.com/library/view/creating-a-data-driven/9781491916902/

Baijens, J., & Helms, R. W. (2019). Developments in knowledge discovery processes and methodologies: anything new?

Cao, L. (2017). Data Science: A Comprehensive Overview. *ACM Computing Surveys*, *50*(3), 1–42. https://doi.org/10.1145/3076253

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. 2000. CRISP-DM 1.0 *Step-by-Step Data Mining Guide.* Technical report, CRISP-DM.

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

Das, M., Cui, R., Campbell, D., Agrawal, G., & Ramnath, R. (2015). Towards methods for systematic research on big data. In IEEE International Conference on Big Data.

Espinosa, J., & Armour, F. (2016). The big data analytics gold rush: a research framework for coordination and governance. In Proceedings of Hawaii International Conference on System Sciences.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "Knowledge Discovery and Data Mining: Towards a Unifying Framework.," *Int Conf on Knowledge Discovery and Data Mining*, pp. 82–88.

Gao, J., Koronios, A., & Selle, S. (2015). Towards a process view on critical success factors in big data analytics projects.

Grady, N. W. 2016. "Knowledge Discovery in Data Science," in 2016 IEEE International Conference on Big Data, pp. 1603–1608.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105.

Hevner, A. R. (2007). *A Three Cycle View of Design Science Research A Three Cycle View of Design Science Research*. *19*(2), 87–92.

Li, Y., Thomas, M. A., Osei-Bryson, & Muata, K. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems*, *91*, 1–12. https://doi.org/10.1016/j.dss.2016.07.003

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. https://doi.org/10.1017/S0269888910000032

Nascimento, D., Santana, G., & de Oliveira, A. A. (2012). *An Agile Knowledge Discovery in Databases Software Process*. (c), 56–64. https://doi.org/10.1007/978-3-642-34679-8_6

Okoli, C., & Schabram, K. (2011). A Guide to Conducting a Systematic Literature Review of Information Systems Research. *Sciences-New York*, *10* (2010). https://doi.org/10.2139/ssrn.1954824

Peffers, 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.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact evaluation in information systems design-science research. *18th Pacific Asia Conference on Information Systems*, 16.

Provost, F. & Fawcet, T. (2013). *Data Science for Business* (First edition ed.): O'Reilly Media Inc.

Saltz, J. S. 2015. "The Need for New Processes, Methodologies and Tools to Support Big Data Teams and Improve Big Data Project Effectiveness," *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 2066–2071.

Saltz, J., & Shamshurin, I. (2015). *Exploring the process of doing data science via an ethnographic study of a media advertising company.* Paper presented at the 2015 IEEE International Conference on Big Data (Big Data).

Saltz, J., & Shamshurin, I. (2016, 5-8 Dec. 2016). *Big data team process methodologies: A literature review and the identification of key factors for a project's success.* Paper presented at the 2016 IEEE International Conference on Big Data (Big Data).

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, 1013–1022.

Saltz, J., Shamshurin, I., and 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), pp. 2720–2728. (https://doi.org/10.1002/asi.23873).

Saltz, J., Wild, D., Hotz, N., & Stirling, K. (2018). *Exploring Project Management Methodologies Used Within Data Science Teams*. in Twenty-Fourth Americas Conference on Information Systems, New Orleans, 2018, pp. 1–5.

Saunders, M., Lewis, P., & Thornhill, A. (2016). *Research Methods For Business Students* (Seventh edition ed.): Pearson Education Limited.

Sharda, R., Delen, D., & Turban, E. (2018). *Business Intelligence, Analytics, and Data Science* (Fourth edition ed.): Pearson Education Limited.

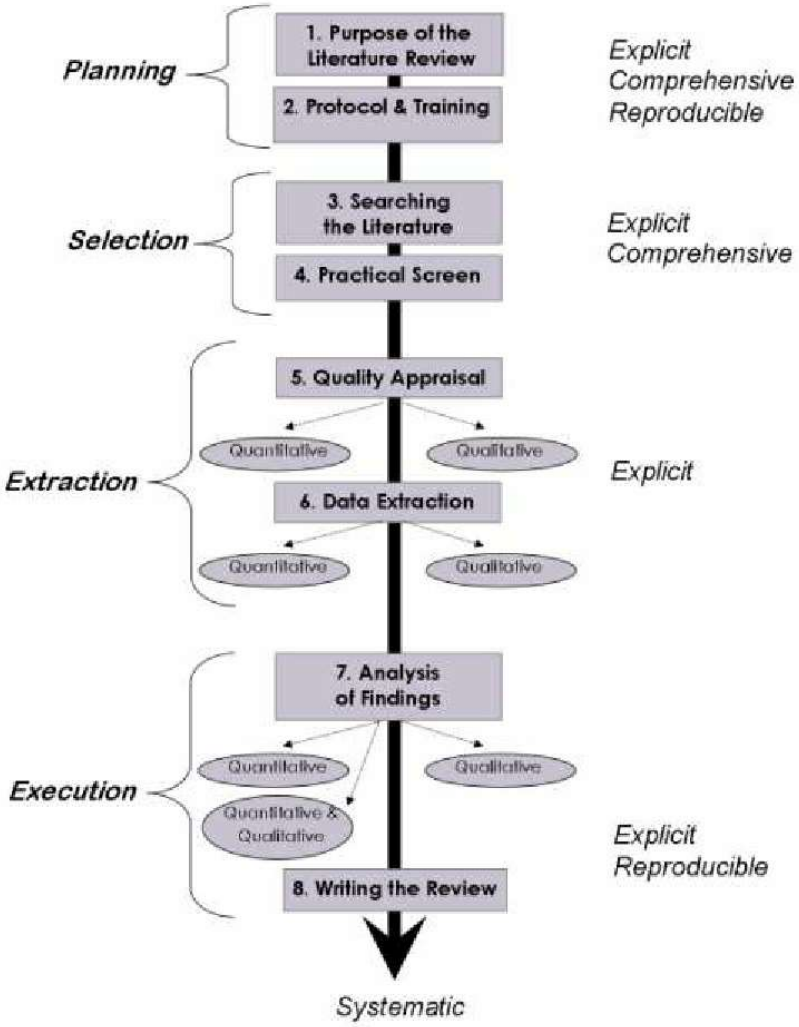
Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, *25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining.* Paper presented at the Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.

Yin, R. K. (2009). *Case study research: Design and methods 4th ed.* Paper presented at the United States: Library of Congress Cataloguing-in-Publication Data.

Bijlage 1 – Literatuurstudie

Als handleiding voor de literatuurstudie is het Systematic Literature Research (SLR) model van Okoli and Schabram (2011) gebruikt. Zij behandelen een concreet stappenplan voor het uitvoeren van een literatuurstudie. (De uitvoering hiervan wordt in de volgende paragraaf uitgewerkt). Onderstaand figuur 1 geeft het model van Okoli and Schabram en de stappen weer:



Figuur 1 Systematic Literature Research (SLR), Okoli and Schabram (2010)

Uitleg te nemen stappen

1. Doel van de literatuurstudie: Wat zijn de beoogde doelen van de literatuurstudie. Dit is noodzakelijk om de beoordeling expliciet te maken voor zijn lezers.
2. Protocol en training: Als er meerdere onderzoekers betrokken zijn, is het van cruciaal belang dat de onderzoekers volledig op de hoogte zijn en instemmen met de gedetailleerde procedure die moet worden gevolgd. Dit vereist zowel een schriftelijk, gedetailleerd

protocoldocument als training voor alle onderzoekers om consistentie in de uitvoering van de review te garanderen.

3. Zoeken naar literatuur: Wat zijn de details van het literatuuronderzoek, en hoe wordt de volledigheid van de zoekopdracht gewaarborgd.
4. Praktisch screening: Screening voor opname. Welke praktische uitsluitingscriteria zijn gebuikt voor het reduceren van het aantal artikelen en hoeveel artikelen zijn vervolgens.
5. Kwaliteitsbeoordeling: Screening op uitsluiting. Welke artikelen zijn van onvoldoende kwaliteit om in de overzichtssynthese te worden opgenomen.
6. Gegevensextractie: Nadat alle artikelen die in de beoordeling moeten worden opgenomen, zijn geïdentificeerd, wordt systematisch de toepasselijke informatie uit elke artikel gehaald.
7. Synthese van studies: Analyse: deze stap omvat het combineren van de feiten uit de artikelen.
8. Schriftelijke beoordeling: Naast de standaardprincipes die moeten worden gevolgd bij het schrijven van onderzoeksartikelen, moet het proces van een systematisch literatuuronderzoek voldoende gedetailleerd worden gerapporteerd dat de resultaten van de beoordeling onafhankelijk kunnen worden gereproduceerd

Deze literatuurstudie probeert antwoord te geven op de onderzoeksvraag en sub-vragen uit hoofdstuk 1.4. De uitkomsten van deze literatuurstudie zullen worden weergegeven in een overzicht om de ontwikkeling van het onderzoeksgebied weer te geven.

Uitwerking van de stappen 1 tot 8 van Okoli and Schabram (2011)

- Doel van de literatuurstudie:
  De literatuurstudie heeft als doel het beantwoorden van de deelvragen uit hoofdstuk 1:
  - Welke projectmethoden of processtappen voor data science projecten kent de literatuur (literatuurstudie)
  - Welke typen data science projecten kent de literatuur/ wat zijn hun karaktristieken? (literatuurstudie)
  - Welke projectstappen kunnen worden gebruikt om tot een juiste projectmethode te komen? (literatuur studie/case studie)
- Protocol:
  Voor het protocol is de Systematic Literature Research (SLR) methode van Okoli and Schabram (2010) gebruikt. Deze methode geeft een concreet stappenplan voor het uitvoeren van een literatuurstudie. Bijkomend voordeel is dat hierdoor het onderzoek beter reproduceerbaar is, wat de validiteit ten goede komt
- Zoeken naar literatuur:
  Allereerst is gestart met een set van 6 artikelen welke werden aangereikt door de begeleiders van de Open Universiteit. Deze set is gebruikt voor de gebiedsverkenning en voor het opstellen van de juiste zoektermen voor de query. Als zoektermen voor de query is gekozen voor de volgende combinaties:

	Zoektermen (AND)		
Zoektermen (OR)	Data Analitycs	Project*	Method*
	Data Science	Process	Model*
	Knowledge Discovery		

Tabel 1, zoektermen

- Als query is gekozen voor TI= ((Data analytics OR Data Science OR knowledge discovery) AND (Project* OR Process))

In eerste instantie is gezocht vanuit de bibliotheek van de Open Universiteit, echter was het daar niet mogelijk de resultaten in 1 overzicht te exporteren voor analyse. Als zoekmachines is in overleg met de begeleider gekozen voor de databases van Web of Science en IEEE in verband met hun affiniteit met het onderwerp. Deze databases werden benaderd vanuit de bibliotheek van de Open Universiteit. Als periode is gekozen voor artikelen van maximaal 9 jaar oud (Web of Science) of vanaf 2010 (IEEE). Dit is na de verschijning van het overzichtsartikel van Mariscal, Marbán, and Fernández, (2010). "A Survey of Data Mining and Knowledge Discovery Process Models and Methodologies,". Hiermee is uitgegaan van de compleetheid van dit artikel, waardoor jongere artikelen een toevoeging op dit artikel zouden zijn. Daarnaast zijn alleen Engels-talige artikelen geselecteerd in verband met de leesbaarheid.

➔ Deze query leverde uiteindelijk bij Web of Science 47 unieke artikelen en IEEE 96 unieke artikelen. Deze 2 lijsten zijn vervolgens praktisch gescreend en de vervolgens zijn de resultaten samengevoegd.

- Praktisch screening:

De eerste praktische screening was op basis van titel. Sluit de titel aan bij het onderwerp van het onderzoek. Dubbele artikelen welke al door de begeleider waren aangereikt en ook in deze lijst staan, zijn hier ook verwijdert uit de selectie. Eventuele twijfelgevallen werden met een ja beantwoord zodat er verder inhoudelijk gescreend kon worden.

De tweede screening was op basis van samenvatting. Sluit de inhoud van de samenvatting aan bij het onderzoek en kan op basis van deze samenvatting worden verwacht dat de inhoud van dit artikel bijdraagt aan het beantwoorden van de onderzoeksvragen?

De derde screening was op basis van inleiding/conclusie en het scannen van de artikelen. Tijdens deze screening zijn de artikelen met een te technische inhoud en de artikelen welke niet specifiek over typen data science projecten, projectmethoden of waardebepaling voor data science projecten gingen afgevallen.

Uiteindelijk resulteerde deze screening in een set van 12 artikelen. Na snowballing door de artikelen kwamen hier nog 10 artikelen bij.

Het proces is weergegeven in onderstaand overzicht:

		Web of Science	IEEE
1	Start	47	96
2	Screening op titel	14	23
3	Screening op samenvatting	7	11
4	Samenvoegen unieke artikelen	17	
5	Screening op inhoud	12	
	Snowballing	10	
	Totaal artikelen	22	

Tabel 1 Resultaten screening

- Kwaliteitsbeoordeling:
  Voor de beoordeling van de kwaliteit is gecontroleerd of de artikelen peer reviewed waren. Daarnaast is ook gecontroleerd of de artikelen uit een als betrouwbaar en kwalitatief geachte database verkregen of beschikbaar zijn. Hier zijn geen artikelen afgevallen.
- Gegevensextractie:
  De uiteindelijke selectie van de artikelen is geheel doorgenomen en de gevonden informatie verwerkt in een mindmap en in samenvattingen van artikelen.
- Synthese van studies:
  Alle gevonden informatie welke relevant is voor de onderzoeksvragen is verwerkt in het theoretisch kader in hoofdstuk 2.3 en daar gestructureerd naar de twee onderwerpen uit de deelvragen.
- Schriftelijke beoordeling:
  De zoekcriteria zoals de gebruikte queries, databases, periode, taal etc. worden specifiek genoemd, zodat het onderzoek kan worden gereproduceerd.

Bijlage 2 – Interview guide

Ronde 1, semi gestructureerde (oriënterende) interview met specialisten

Doel is om inzicht te krijgen in de ervaring van de geïnterviewde en de organisatie over het gebruik van project methoden voor data science projecten.

Voorafgaande contact met de geïnterviewden:

- Afstemmen wat het doel is van de studie en het interview.
    - Afstemmen dat er een transciptie van het gesprek wordt gemaakt
    - Afstemmen dat het gesprek wordt opgenomen voor de transcriptie
    - Afstemmen dat het gesprek wordt geanonimiseerd

De interviewvragen zijn gebaseerd op de onderwerpen van theoretisch framework uit hoofdstuk 2.3.3. Mogelijke vragen/gespreksonderwerpen:

- Kunt u zich introduceren? (om af te stemmen wat hun rol m.b.t. Data science projecten is)
- Wat is uw ervaring met / kennis over verschillende typen data science projecten? (bespreek hier eventueel de verschillende projecteigenschappen)
- Wat is uw ervaring met / kennis over verschillende projectmethoden? (bespreek hier eventueel de karakteristieken)
- Maken jullie gebruik van een (vast) stappenplan?
    - Welke stappen bevat deze?
    - Indien onbekend, beschrijf het proces van een nieuwe data analyse opdracht

Na de eerste ronde wordt de verkregen informatie verwerkt en toegepast op het ontwikkelde artefact. Kijkend naar de projectkarakteristieken van de projecten binnen de case organisatie, wat zouden dan logische processtappen zijn?

Deelnemers/specialisten:

- Director - Global Research & Analytics
- Content Marketing Specialist - Marketing Research & Analytics
- Business Manager - Customer Service
- Project Engineer - Research & Development (als analist van de afdeling R&D)
- Product Owner - Data & Metrics

Ronde 2, eerste focusgroep interview

Na de eerste ronde interviews was er nog onvoldoende informatie beschikbaar als input voor de evaluatie/creatie van het framework, daarom is een tweede ronde in de vorm van een focusgroep interview gehouden. Doel is om een dieper inzicht te krijgen in de professionele mening van de specialisten m.b.t. het framework. Hoe verhouden de projectkarakteristieken zich t.o.v. de processtappen.

Note, de eerste en tweede ronde dienen (samen met het documentonderzoek) als input voor de formatieve evaluatie.

Het voordeel van een focusgroep interview is dat de geïnterviewden samen kunnen discussiëren over de vragen en antwoorden. Hierdoor kunnen ze ook elkaar bevragen en daarmee tot een gezamenlijk volledig(er) antwoord komen (Saunders, Lewis, & Thornhill, 2016).

Introductie als start van focusgroep interview:

- Afstemmen wat het doel is van de studie en het interview.
    - Afstemmen dat er een transciptie van het gesprek wordt gemaakt
    - Afstemmen dat het gesprek wordt opgenomen voor de transcriptie
    - Afstemmen dat het gesprek wordt geanonimiseerd
- Uitleg over processtappen van CRISP-DM en toegevoegde thema's
- Uitleg over projectkarakteristieken

De interviewvragen/ gespreksonderwerpen zijn gebaseerd op de informatie uit de eerste ronde en hier wordt dieper ingegaan op het artefact uit hoofdstuk 2.3.3. Mogelijke vragen/ gespreksonderwerpen:

- Heeft het type project invloed op de te gebruiken processtappen?
    - Zo ja, welke processtappen zijn dan belangrijk
- Zijn er binnen de analytical context factoren welke invloed hebben op de te gebruiken processtappen?
    - Zo ja, welke processtappen zijn dan belangrijk
- Zijn er binnen de data context factoren welke invloed hebben op de te gebruiken processtappen?
    - Zo ja, welke processtappen zijn dan belangrijk
- Zijn er binnen de team context factoren welke invloed hebben op de te gebruiken processtappen?
    - Zo ja, welke processtappen zijn dan belangrijk
- Zijn er binnen de organizational context factoren welke invloed hebben op de te gebruiken processtappen?
    - Zo ja, welke processtappen zijn dan belangrijk
- Gebruikt u processtappen welke nog niet benoemd?
    - Zo ja, welke?
- Herkend u projecteigenschappen welke nog niet zijn genoemd?
    - Zo ja, welke?

Deelnemers/specialisten

- Project Engineer - Research & Development (als analist van de afdeling R&D)
- Data Engineer - Data & Metrics
- Global Busines Analist - Sales & Marketing

Ronde 3, tweede focusgroep interview

Doel van de derde ronde is om de toepassing op het ontwikkelde artefact te bespreken en evalueren met de specialisten.

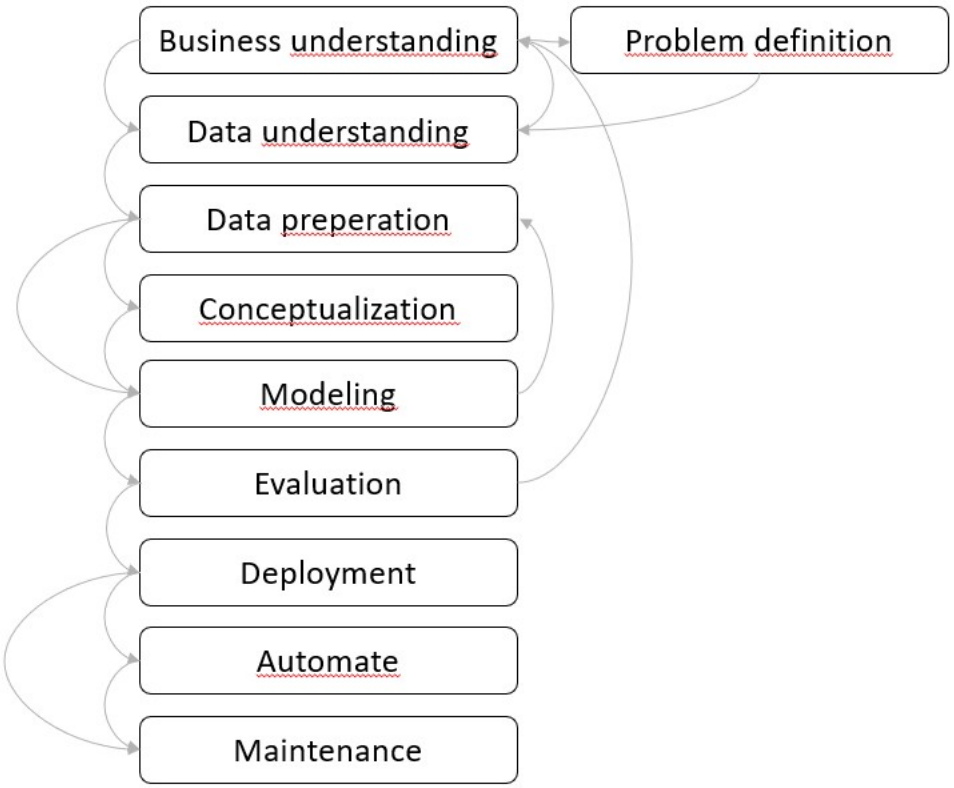
De interviewvragen/ -onderwerpen zijn gebaseerd op de informatie uit de eerste en tweede ronde en op het framework. Tevens is er gebruik gemaakt van onderstaand discussiemodel met processtappen als basis voor de verificatie. Dit discussiemodel is gebaseerd op de processtappen van CRISP-DM, plus de extra processtappen welke zijn gevonden in de literatuurstudie. Tijdens deze sessie zijn de extra proces stappen t.o.v. CRISP-DM besproken (en hun toegevoegde waarde), plus welke proces stappen bij welke project karakteristieken passen. Doel van de derde ronde is om de toepassing op het ontwikkelde raamwerk te toetsen met de geïnterviewden en als basis voor de summatieve evaluatie. Tijdens dit focusgroep interview wordt het discussiemodel getoetst op de vijf criteria van Prat et al. (2014):

- Goal
- Environment
- Structure
- Activity
- Evolution

Mogelijke vragen/ gespreksonderwerpen:

- Was het discussiemodel duidelijk?
    - Zijn er nog zaken welke moeten worden toegelicht?
- Is het model uws inziens relevant en praktisch toepasbaar?
    - Zijn er nog aanvullende opmerkingen?
- Is het model uws inziens compleet?
    - Zijn er nog eventuele aanvullingen?
- Wat is uw uiteindelijke reflectie over het model?

Discussiemodel



Deelnemers/specialisten

- Project Engineer - Research & Development (als analist van de afdeling R&D)
- Data Engineer - Data & Metrics
- Global Busines Analist - Sales & Marketing

Bijlage 3 – Codeerschema

111 - Project Methods - CRISP-DM - Business Understanding
112 - Project Methods - CRISP-DM - Data Understanding
113 - Project Methods - CRISP-DM - Data Preperation
114 - Project Methods - CRISP-DM - Modeling
115 - Project Methods - CRISP-DM - Evaluation
116 - Project Methods - CRISP-DM - Deployment
117 - Project Methods - CRISP-DM - Waterval Process
121 - Project Methods - Aanvullende Thema's - Life cycle selection
122 - Project Methods - Aanvullende Thema's - Conceptualization
123 - Project Methods - Aanvullende Thema's - Automate
124 - Project Methods - Aanvullende Thema's - Support / Maintanance
131 - Project Methods - Aangepast Process - Hybride Waterval + Agile
132 - Project Methods - Aangepast Process - Agile Scrum
133 - Project Methods - Aangepast Process - Agile Kanban
212 - Project Characteristics - Projecttype - Eenmalig
212 - Project Characteristics - Projecttype - Routinematig
221 - Project Characteristics - Data Context - Variety
222 - Project Characteristics - Data Context - Volume
223 - Project Characteristics - Data Context - Velocity
224 - Project Characteristics - Data Context - Veracity
231 - Project Characteristics - Analytical Context - Type of Analysis
232 - Project Characteristics - Analytical Context - Compute Intensity
233 - Project Characteristics - Analytical Context - Criticality of Timelines
241 - Project Characteristics - Team Context - Size of Team
242 - Project Characteristics - Team Context - Virtuality of Team
243 - Project Characteristics - Team Context - Manager Experience
251 - Project Characteristics - Organizational Context - Size of Organization
252 - Project Characteristics - Organizational Context - Organization Culture & Process
253 - Project Characteristics - Organizational Context - Organization Culture & ROI
254 - Project Characteristics - Organizational Context - Number of Data Science Teams