Applying data science to improve municipal youth care

Master thesis

Student:

Floris Smit (s1253999)
Business & IT

Supervisors:

K. Sikkel
C. Amrit

UNIVERSITEIT TWENTE.

M. Imrich

Xomnia

Management summary

Purpose

Data science has transformed many businesses and organisations. Government organisations see the benefits of applying data science but do not have the required experience to reap the benefits. The aim of this research is to find out which challenges municipalities face when applying data science.

Method

In this research, a data science project in a municipality has been studied in a case study. Five stakeholders from different positions within the organisation have been interviewed to provide insights in the challenges of applying data science from multiple perspectives. The interview design is based on a systematic literature review and a preliminary study of data science in practice.

Results

The most important challenges found in the case study are:

- Creating a useful research question was difficult.
- Data quality and quantity was low.
- There was uncertainty around privacy laws.

These challenges could apply to many other data science projects in government organizations.

Privacy

Privacy of citizens is very important to municipalities so it makes sense that they are conservative regarding this subject. New privacy laws cause a lot of uncertainty because it is not clear what is allowed and what is not allowed. This causes a paralyzing fear which slows down data science projects throughout the municipality. The privacy issues experienced in the case study have to be solved before data science can be successfully applied.

Conclusions

Based on the results of this case study, one can conclude that data science puts high demands on the data and the stakeholders involved. These demands could be translated to analytical maturity and data management maturity.
Analytical maturity describes to what extent the organisation uses data-driven decisionmaking. In the case study, some stakeholders had little analytical experience, which lead to the difficulty to find good research questions. Having a higher analytics maturity in the organisations should familiarise stakeholders with analytical thinking.
Data management maturity describes how well an organisation can turn their data into an asset and includes practices ranging from the strategic to the infrastructure level. Data quality and quantity are problems that were encountered in the case study that indicate a low level of data management maturity.

Organisations starting with data science should check their analytical and data management maturity levels. A minimum level of data management maturity is needed before a data science project can be successful. Higher data management maturity provides data for data scientists to use in their experiments. Higher analytics maturity prepares the organisation for data science by introducing basic data-driven decisionmaking.

Recommendations

Based on the case study, the following recommendations could be made to government organisations who are just starting out with data science projects:

- Reduce uncertainty on privacy laws by mapping the grey area between what is and what is not allowed.
- Create showcase projects to educate the organisation
- Provide training on basic data science concepts for business users involved in data science projects
- Stimulate data-driven decisionmaking in the entire organisation (self-service business Intelligence)

Table of contents

1) Introduction

In the past years big data analytics and data science have been used by companies to provide competitive advantages and to provide a unique user experience. This includes providing personalised advertisements, product recommendations and personalised search. Data science technologies continue to develop and mature, causing them to become accessible to more traditional companies. The traditional companies that want to benefit from data science technologies are different from the web based companies where data science started. They have a lot of processes, cultures and paradigms that might not always let them apply data science in a successful way.

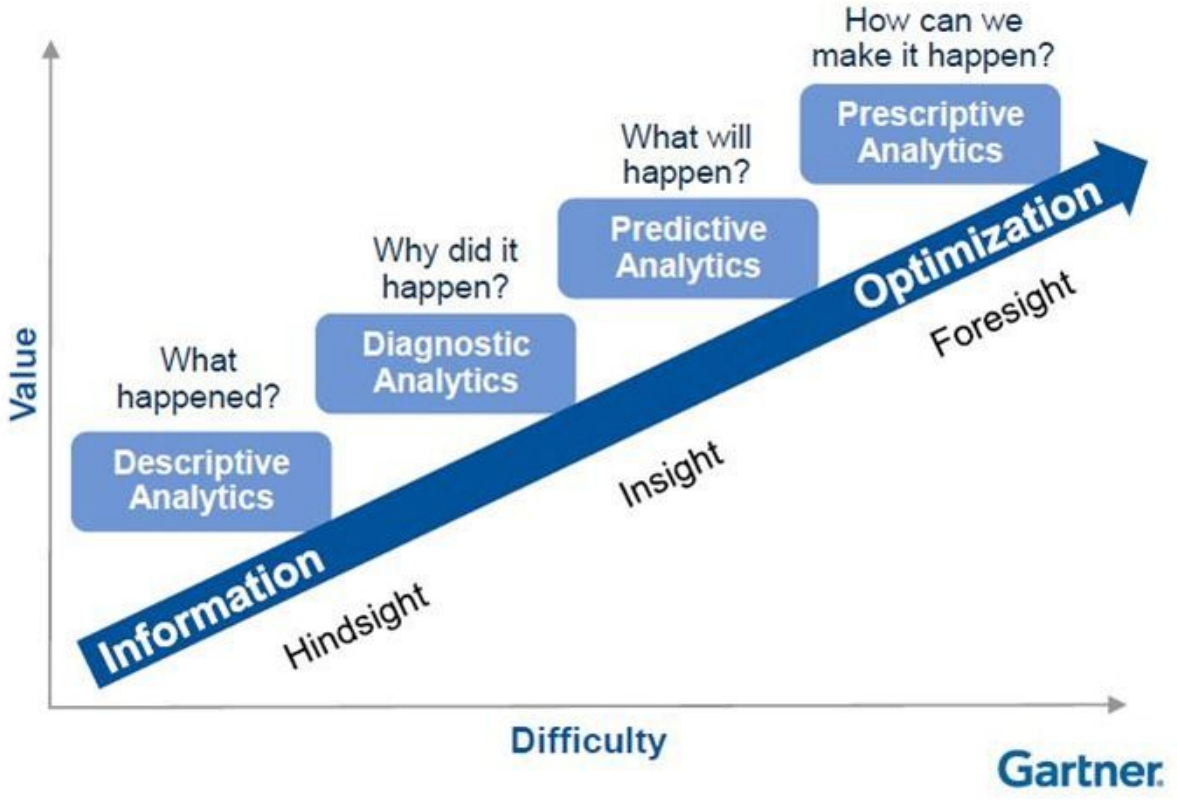

Figure 1. Gartner's analytics maturity model

Organisations have different levels of maturity with regards to their analytics capability. Gartner published the Analytics Maturity Model in 2013 described by the image in figure 1 (Maoz, 2013). It states that organisations should first get familiar with looking back in time before looking forward. Data-driven decisionmaking in organisations requires both IT infrastructure and an analytical mindset of employees. This makes improving the data-driven decisionmaking complex both in terms of IT infrastructure and change management. Large organisations that aspire to do data science will get more value when they are already successfully using business intelligence.

A typical data science project usually start with a business problem or need. These are often focussed on increasing revenue, reducing costs or mitigating risks. A data scientists talks with stakeholders in the organization to find relevant topics. The data scientist explores the

available data to look for inconsistencies and to test hypotheses and assumptions of the stakeholders. This will improve the domain knowledge and understanding of the available data. The data scientist can pick a relevant problem to solve based on the needs of stakeholders, data available and complexity of the problem. The data scientist prepares the data and creates a model to predict something that helps in solving the problem. This model is continuously improved and enriched with more data based on feedback of the stakeholders. When the performance of the model is good enough, it can be deployed to a production environment where the organization can use the model. This can be in the form of predicted values embedded in a dashboard. The model can also be deployed as an API to integrate with an existing application. After a model is integrated in existing IT, employees can be supported in their normal working environment and some decisions can even be automated.

Municipalities and government organisations are traditionally organisations that lag behind in technological developments. They adopt technologies when they are getting older, and are hesitant to adopt new, unproven technologies. Big data analytics and data science follow the same rules, but the municipalities willingness to experiment with them is also increasing. While selling ads and products is useful for companies the benefits to society are limited. Municipalities and governments start to recognise the possible benefits these technologies can give to society, increasing their willingness to initiate these projects. Being able to predict certain aspects that are relevant to the organisation can help employees make better decisions that benefit society. At the same time, they are also conservative in using the data of citizens. They know their actions and motives should be transparent to citizens, and not all citizens like the idea that the government uses their data.
On top of that IT projects in general are a difficult topic for government organisations, as one third of the projects fail to meet their requirements either in time or within budget. Governmental organisations are often very bureaucratic and implementing projects methods like agile software development is often not in line with the organisational culture.
These factors combined create an environment where applying data science could be quite challenging.
This research is a case study which aims to identify the challenges government organisations face when applying data science. There is research done on challenges in data science but not many case studies. The scientific contribution of this research will be the validation of the challenges found in literature in practice. These insights can also be used by organisations to mitigate challenges in their own projects. The contribution to society and practice will be money saved on government projects as well as the benefits of data science to society.

2) Research design

2.1) Problem statement

Data science and big data are new techniques that could provide benefits to society when applied in municipalities and other government organisations. Municipalities have a lot of data but lack the knowledge to use it. There has not been much research on applying data science in municipalities.
This research will be an observational case study in a municipality in the Netherlands which we will call King's Landing throughout the paper. This is done to ensure that stakeholders can speak their mind without risking negative publicity. In this case the municipality will start a data science project to improve youth care. The goal of the project is to explore the benefits data science can have for them, as well as training their employees in new technologies.

2.2) Research objectives

The goal of this research is to study opportunities and challenges in applying data science within municipalities. This will help other municipalities and similar organisations avoid the same pitfalls when doing data science projects. These findings can be used to make statements about data science in general.

2.3) Research questions

The research consists of the following research questions:

Q1 What are the challenges and opportunities of doing data science projects according to literature?
Q2 What are challenges of applying data science in practice?
Q3 What challenges are expected to be important when doing data science projects in municipalities?
Q4 How does King's Landing address these challenges?
Q5 What can we learn from King's Landing's experiences?

The result of each question can be used in answering the next question. The dependence structure of the research questions and their deliverables will be elaborated in the next chapter.

2.4) Research methodology

The research is exploratory and has been structured using the schema in figure 2 Each block describes a deliverable that answers a research question. Arrows and accolades between deliverables represent a dependency relation. Deliverables are used as input to create the next deliverables. By using this approach the end results are grounded in literature and practice, and the reasoning behind the research structure becomes transparent.

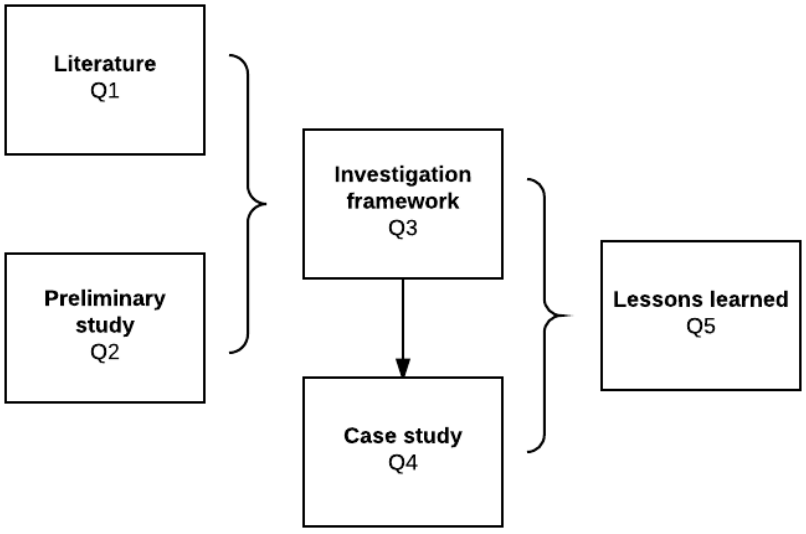

Figure 2. Research dependency structure

Each of the deliverables make use different research methods. The grounding in literature (Q1) has been done using a systematic literature study. The literature was consolidated in a concept matrix as defined by Webster & Watson (2002). The preliminary study (Q2) makes use of both a literature study and interviews. The investigation framework (Q3) makes use of CRISP-DM (Chapman, 2000) to structure the interview design. During the case study (Q4) at King's Landing interviews were done to collect data.
Each of the research methods will be described in more detail in their respective chapters.

2.5) Research approach

The literature study has been done to find challenges and opportunities in applying data science that have been described in literature. This knowledge is combined with a preliminary problem analysis. The problem analysis consists of interviews that have been conducted at Xomnia, the consulting company that is involved in the case study. Several challenges found in the problem analysis also apply to this research.
From these two studies an investigation framework has been defined. This framework will be a list of challenges that can originate in either literature and the research topics.
Based on this investigation framework the case study was designed. Several relevant challenges were selected and investigated in practice by interviewing 5 stakeholders of the project. During these interviews the interviewees were asked about the challenges they have experienced in this project. Based on the the interviews it will become clear which challenges municipalities face when applying data science.
The results of the case study have been used to extract lessons learned about challenges in applying data science. Some of these lessons and findings can be generalised to other organisations.

3) Literature study

In this chapter the literature related to the research questions will be presented. Since data science is a new term there is a lot of discussion on the definition of data science. The first part of this literature study will be about the definitions of data science found in literature. After that CRISP-DM will be explained because is used in the case study project, and it will be used to structure parts of the interviews.
Data science is a new concept but companies are getting familiar with it, and they know they need data scientists to get more value out of their data. The opportunities these companies see in data science will be discussed next. Companies that have just started creating data science teams are encountering challenges. These challenges will be described and categorized into concepts.

3.1) Literature strategy

The literature was found by querying Scopus, Web of Science and Google scholar. The queries used where: “data science challenges”, “data science challenges projects”. This resulted in a collection of papers which were not all relevant. Based on the title and abstract a first selection was made. Papers and books which had a purely technical focus were excluded. The resulting papers were read and searched for explicit challenges. These challenges categorised in concepts. The choice of these concepts was based on the nature of the challenges. While different categorisations are possible a choice was made that would support the case study.

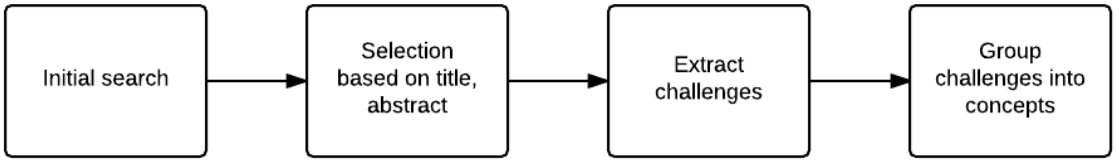

Figure 3. Literature strategy

3.2) Defining data science

There are several definitions of data science found in literature. Harris & Mehrotra (2014) define data science as “an emerging profession that leverages programming and statistical skills to solve business problems.”

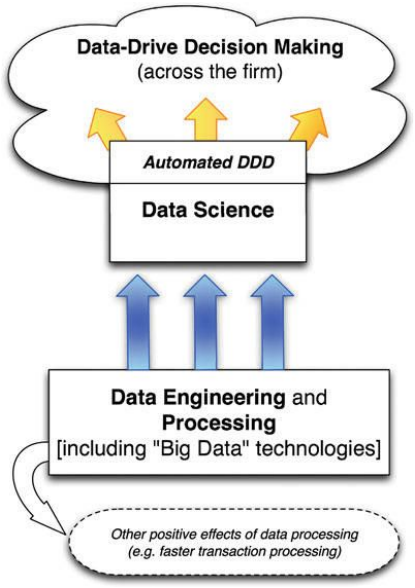

Figure 4. Data science definition by Provost and Fawcett (2013)

Provost and Fawcett (2013) have a more nuanced definition of data science. They argue that it has been hard to define what data science is, because it is intertwined with other areas like big data and data-driven decisionmaking. They study the relation with these other areas, and by doing this they identify what the fundamental principles of data science are. Data science is more than just data mining, as a data scientist is able to look at business problems from a data perspective. Data science uses data engineering and data processing principles to improve (automated) decision making (see figure 4). Big data technologies fall under data engineering and data processing, and are mostly used to support data science. Provost defines data science as “A set of fundamental principles that support and guide the principled extraction of information and knowledge from data”.

3.3) CRISP-DM

Next to a definition it is also useful to understand the activities involved in doing data science. One of the most widely accepted methodologies for doing data mining is the Cross Industry Process for Data Mining (CRISP-DM). (Chapman 2000) This method describes how to do data mining projects by dividing the work involved in different phases (see figure 5). It is a widely accepted methodology that emerged from the combined knowledge of the leading industry. There are other methodologies in the industry like KDD and SEMMA. SEMMA lacks the deployment and business understanding equivalent and KDD has similar phases. (Azevedo & Santos, 2008) In the case study project CRISP-DM is used instead of SEMMA or KDD. Therefore this chapter will focus on CRISP-DM.

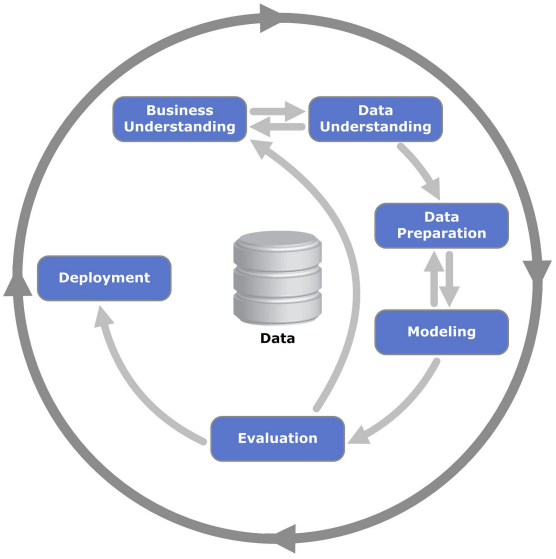

Figure 5. CRISP-DM cycle

Business understanding
When doing Data Science it is important to understand what the business needs. In some cases the core business is complex, with a lot of specific terminology. It is hard to make a model that is useful for the business without business understanding.
Data understanding
The business uses a number of different IT systems, which all generate data. It is important to understand what the data means. It is also useful to find out if there are any data quality problems.
Data preparation
During the data preparation phase data is combined with a different dataset which is suitable for analysis. Data from multiple sources has to be combined, cleaned and enriched.
Modeling
During the modeling phase the prepared data is fed to different algorithms. These different algorithms can be compared on performance. The best algorithms are selected and improved.
Evaluation
When a model is developed it is time to evaluate it within the organisation. Before deploying the model it is useful to know if the model still supports business goals.
Deployment
When the model's performance is sufficient, and it is evaluated in the organisation it can be deployed. The deployment can be a report, a presentation or the automation of a process.

Dividing the data mining process in these phases is quite useful. It provides Data Scientist with a framework to place their process in perspective, and a way to decouple the process. It also gives other stakeholders in the organisation a way to understand on a basic level how data mining works.

3.4) Challenges & Opportunities

The literature that has challenges and opportunities will be summarised by using a concept matrix as defined by Webster & Watson (2002). This will give a high level overview of what challenges and opportunities are found in literature. It will provide insight in which concepts are covered by a lot of papers and which papers cover a lot of concepts. See table 1 for the concept matrix that followed the literature study of this thesis.

	Opportunities			Challenges				
	Measure things in greater detail	**Data driven decisionmaking**		**Starting data science projects**	**Data science team dynamics**	**Company mindset**	**Data science research methods**	**Privacy & ethics**
Cao (2016)				X	X	X		
Rose (2016)				X	X	X	X	
McAfee (2012)				X	X	X		
Viaene (2013)					X			
Drew (2016)								X
Carter & Sholler (2016)							X	X
Provost & Fawcett (2013)		X				X		
Brynjolfsson (2011)		X						
Khan (2013)	X							
Swan (2013)	X							

Table 1: Concept matrix with challenges and opportunities found in literature.

3.4.1) Opportunities: how can companies profit from Data Science

According to literature there are a number of opportunities for organisations applying data science.

- Measuring things in greater detail
- (Increased) data-driven decisionmaking

Measure things in greater detail

The exponential growth of computational power and storage predicted by Moore is still true today. At the same time the costs of electronics is decreasing, making way for internet of

things (IoT) applications. These applications allow companies to bridge the gap between the physical world and the digital domain. All these applications generate large amounts of data, which can be analyzed using data science.

Swan (2013) introduces the quantified self as a development that is enabled by the development of new technologies. It allows for the continuous monitoring of behaviour and biological processes of individuals. This development brings a whole range of new applications that all generate data that needs to be analyzed using big data technologies and data science techniques. The difference with traditional methods is that the quantified self allows for continuously monitoring these individual aspects, while traditional methods rely on periodic samples and surveys. One of the opportunities highlighted by these authors is the ability to measure behavioural, environmental, biological and physical aspects of individuals providing the basis for new insights.
Another field that uses data science and big data technologies is smart cities. In a case study by Kahn (2013) a big data architecture that can support all the sensors and other devices is designed. The goal of smart cities is to use IoT technologies to solve urban challenges. IoT devices allow for a wide variety of aspects of the city to be measured in greater detail. An example of such an application would be smart parking. IoT devices measure all major parking places in the city giving a real time overview of parking place occupancy. This information can be used to direct traffic to available parking spaces in real time. A data scientist could analyse all this data to provide policymakers with the best location for a new parking place.

Data driven decisionmaking
A concept related to the last is improved data-driven decisionmaking. Decision making is mostly done based on experience and the gut feeling of managers. When more data about the business and the environment becomes available, it opens up ways for managers to base decisions on data instead of gut feeling. The reasoning behind this is that decisions based on facts are better decisions than decisions based on gut feeling and experience.

Provost & Fawcett (2013) argue that data science should support data-driven decisionmaking. They identify two types of decisions. The first being decisions about general strategy that require discoveries. The other type is small, repeated decisions which occur very often. When applying data science to these smaller repeatable decisions a small improvement in the decision making process already has effect because of the scale. These decisions are also candidates for automating the decision making process instead of supporting it. A prime example of this is the automated placement of advertisement since the decision process is automated. Split second decisions are made to show a specific advertisement to a specific user based on the past behaviour of that user.

Brynjolfsson (2011) has done empirical research to the effectiveness of data-driven decisionmaking. They used survey data to conclude that use of data-driven decisionmaking gives a 5-6% advantage in output and productivity over companies that don't use it.
In a case study by Drew (2016) in government organisations and claims that data science can be used to improve data-driven decisionmaking. He also identified principles regarding

the ethical aspects of data science that governments should abide. Using data to improve or automate decision making in a government context is a controversial topic.

3.4.2) Challenges in applying data science

Challenges in starting data science projects

The first challenges when doing data science arise when the organisation is preparing to start the projects. These challenges often come from a lack of knowledge in the organisations, or wrong expectations about what data science needs to be successful.
I will discuss the following challenges:

- Pitfalls about data science concepts (Cao, 2016)
- Overfocus on technology(Rose, 2016)
- Talent management (McAfee, 2012)
- Technology gap with existing IT (McAfee, 2012)
- Data volume and infrastructure pitfalls (Cao, 2016)

Cao (2016) argues that there can be misunderstanding within an organisation about what data science is. A big part of data science has roots in statistics, so some people might question the need for a new concept. Other fields data science has roots in are data engineering, information sciences and data analysis. There might be people in the organisation that claim that data science doesn't offer anything new. While data science is largely based on these other fields it does offer a new multidisciplinary way of working to solve complex problems based on large volumes of data.
Rose (2016) argues that one of the key challenges is that companies tend to focus on the hardware and technology part of a data science project. They believe that if they build a cluster and collect all their data, insight and knowledge will automatically follow. There are numerous cases where organisation have built a big cluster to gain insights, but when it is finished they don't know what to do with it, or they lack the knowledge in the organisation. There is little thought about what to gain from the project, and how insights can be generated.
Rose argues that data science is about exploration and that data is not the product, insight is. Having a flexible data science teams with ad hoc solutions can be just as effective as having a large infrastructure. Effective data science teams can be messy as they will use a variety of different tools.
It is clear that data science relies a great deal on having a good team that has the right knowledge. McAfee (2012) says that talent management of data scientists is one of the challenges organisations face. A data scientist has the skills to work with large amounts of data, which are traditionally not taught in statistics classes. They also must speak the language of the business to help leaders transform their business to maximise the advantages of big data. Because the cost of data is dropping these people are in high demand, making it hard for organisations to acquire the talent they need.
Another important group of stakeholders McAfee (2012) highlights is the traditional IT within organisations. The knowledge that traditional IT departments have often does not include big data technologies. But data scientists are dependant on IT to maintain their applications, because they usually lack the skillset required to do this themselves.

The infrastructure required for data scientists to do their work is highly dependant on the data volume. Cao (2016) recognises challenges in the data volume and infrastructure. Often organisations don't know how big their data will be. Cao argues that not only the volume but also the complexity of the data is an important factor to decide whether data science is required to tackle a problem. The infrastructure required to do analysis is more dependant on the volume of data, although Cao argues that organisations can already do a lot of the big data analysis without acquiring the infrastructure.

There are a lot of challenges relating to starting up data science projects found in literature. They can be summarized as follows:

- No clear business goal
- No focus on data science team
- Overfocus on infrastructure and technology

Data science in an organisation should start with a global business goal. Collecting data does not magically give insights. A good data science team will guide the organisation in further specifying the goals and combine the input from stakeholders with knowledge about the data to think of applications that will benefit the organisation.
The well functioning data science team is critical to the success of data science within the organisation. Organisations starting with data science should create a team with a different backgrounds to complement strengths and weaknesses of each data scientist. Some might have more business affiliation, others have a strong passion for math and statistics, while others might have a computer science background. There will be more attention to data science team dynamics in the next chapter.
The infrastructure and tools required should be largely determined by the team. When starting out a relational database to query with SQL could be enough to help the data science team get insight in the data and business. Each data scientist can use their own favorite tools and in the exploratory phases they should have the freedom to use what they want. When the need arises from the team to have dedicated infrastructure a cluster can be created. What is missing from the literature found is use of cloud providers. A lot of technology companies take the lean approach to infrastructure, using cloud providers like Amazon AWS, Google Cloud Platform or Azure to quickly adapt the infrastructure to the organisations need. A key concept in this strategy is to keep all the data in simple storage like Amazon S3 and only create a cluster for analysis when it is required.

Data science team dynamics

As mentioned in the previous section creating a data science team is important when starting data science. When a company becomes more mature the team will grow and the success is still dependant on team interactions. This makes focussing on the dynamics of data science teams an important factor. There are a number of challenges that are identified in literature regarding data science team dynamics.

- Pitfalls about roles and capabilities (Cao, 2016)
- Nurture Versatile Employees (Viaene, 2013)
- Reaching consensus quickly vs. wandering (Rose, 2016)
- Balance between sprints and exploration (Rose, 2016)
- Short experimentation cycle (Rose, 2016)

- Produce Working Solutions (Viaene, 2013)
- “Not invented here”-syndrome (McAfee, 2012)

Cao (2016) states that a lot of people have started calling themselves data scientists, but they actually do data engineering and descriptive analytics. The skill set associated with data science is large, but data scientists who are experts on all are very rare if not non existent. Having a diverse team is important for the performance of the team because it will ensure that all skills required are present in the team.

Viaene (2013) agrees with this and states that data scientists aren't superheroes. They have to engage in conversation with others to produce solutions that benefit the organisation. Therefore it is good to nurture versatile employees, as the stakeholders and domain expert are also very versatile. The versatility of the individuals will help the entire team in communication with the different stakeholders.

Rose (2016) mentions a good indicator for data scientist team dynamics: reaching consensus quickly vs wandering. When in the exploration phase it is important that there is an open culture which allows people to be critical and argue about interpretation. If a team reaches consensus quickly there is a risk they are sharing an assumption that is wrong or they are showing socially polite behaviour. It is important that the team is small enough so everyone feels comfortable disagreeing. When a team is reaching consensus quickly an option is to add someone to the team who is prone to ask questions.

The opposite of reaching consensus quickly is wandering. When the team constantly questions everything and does not reach consensus there is a risk that they spend too much time on the wrong question. Another form of this is when the team is busy answering detailed questions, focussing on the data and failing to see the bigger picture. It is the responsibility of the research lead to ensure that this does not happen, and having a small diverse team helps. Rose states that a good data science team typically consists of two data scientists, a project lead and a research lead.

In the day to day work of the data science team Rose (2016) identifies another delicate balance. Often the data science team gets requests from the organisation which are added to the question board and count toward the sprint. This work usually does not generate new insights as they are specific for a user. So sometimes a data scientist sees something in the data of business and is triggered to explore it. When a data science team allocates all their time to answering questions for the business they won't have time to explore. While exploration gives no direct business value it is very important in gaining high level business understanding and finding new opportunities.

Presenting findings to the business will allow data scientists to validate the models and increase their business understanding. Rose (2016) argues that it is beneficial to work in short experimentation cycles with a presentation and feedback moment after each cycle. This feedback can change the direction of the research, keeping the data science team on the right track. Having feedback at a high intervall will minimise time spent on dead ends.

Something that should also help data science teams is the focus on producing working solution. Viaene (2013) states that insights are not the only result of a data science team, and that the focus should also be on producing working solutions. While insights are valuable working solutions can be implemented in the organisation providing a long term, repeatable benefit.

The organisation has to be flexible enough to accept the insights and solutions. McAfee (2012) warns for the “Not invented here”-syndrome where people in the organisation don't

easily accept new solutions. He argues to stimulate cross functional cooperation by putting the people who understand the problems together with data scientists to produce solutions that benefit the organisation.

Team dynamics are important when applying data science. The following aspect summarize the literature found on this topic:

- Team diversity
- Team methods
- Interaction with the business.

Several times literature mentions the importance to have a diverse team with different backgrounds, skills and competencies. The versatility in the team will help the team interacting with stakeholders. It is good to keep a balance between wandering and reaching consent within the team.
Having regular interaction with the business is a key aspect in the team dynamics. Data science teams should involve domain experts, producing insights and working solutions for the business.

Company mindset

Most companies work in a more traditional way and are turning to data science to become more agile and data driven. For data science to be successful the companies must also change their mindset regarding some of the traditional methods. There are a number of challenges found in literature that are associated with the company mindset.

- Working without objectives (Rose, 2016)
- Leadership (McAfee, 2012)
- Muting HiPPo (Highest paid person's opinion) (McAfee, 2012)
- Data-Analytic Thinking (Provost & Fawcett, 2013)
- Analytics pitfalls (Cao, 2016)

Rose (2016) states that organisational change comes with properly applying data science. Data and data scientists are not static resources that need to be strictly controlled by the organisation. That doesnt mean data scientists don't need input from the organisation as they need direction and domain knowledge. Traditional project management methods rely on strictly defining the objectives and milestones. Managers applying these methods often focus on compliance and management of the team. As data science teams gain new insights they can decide to take a different approach or change the short term goals. Having strictly defined goals will hinder discovery and exploration. When working in a traditional organisation a key challenge is to manage the expectations of other managers. They have successfully used these methods for years, and data science teams need their cooperation to increase business understanding and find new areas to explore.
McAfee (2012) argues that companies that want to be successful need leadership that sets clear goals, defines success and ask the right questions. Although this might seem conflicting with Rose’s argument McAfee’s statements are about the strategic goals of the whole company. McAfee also talks about a new culture of decision making which is more data driven.
Traditionally when there is little data and an important decision has to be made high management will use their intuition to make it. But when more data becomes available it

would be best to sometimes mute the Highest paid person's opinion (HiPPo). Most of them are willing to trust the data when their own intuition disagrees but McAfee believes that too many people rely only on experiences when making decisions.
Making decisions based on data instead of experience also requires knowledge. Provost & Fawcett (2013) state that making data-driven decisions require data analytic thinking capacity. This capacity should be present throughout the organisation for data science projects to succeed. Managers should still be able to communicate with data science teams about projects. When a consultant or data scientist asks input from the organisation the stakeholders involved should have the capacity to assess the proposal. When the entire organisation lacks this capacity the organisation will not really understand what is going on in the data science projects. This will hurt the interaction between the data scientists and the decision makers which can lead to decision makers making the wrong decisions.
Cao (2016) also thinks analytical thinking capacity is crucial and identifies analytical pitfalls. An analyst may present statistically significant findings or summarised data mining results to the business. Then it turns out the business are not interested in these findings. The focus is on the analytics and the data instead of actionable insights and decision making support.

There are a number of authors that mention the change of company mindset in relation to data science. Managing data science teams requires a different approach than traditional managers know. Less focus on strict objectives and control, more focus on guiding and facilitating the team. The mindset of managers throughout the organisation has to change for the organisation to become more data driven. When intuition and the data are conflicting they have to be able to silence their intuition instead of blindly following their intuition. This is easier said than done since most managers have learned to trust their intuition over the years.
When applying data science in a traditional organisation it is important to remember a few things. Managers will be used to working with strict project management methods so it is good to manage their expectations, and communicate about the way the team works. Doing this will improve communications and will slowly change the mindset of those managers. When interacting with decision makers it is important to understand their problems and their view on the organisation. Then the data science team can tempt them to become more data-driven by offering solutions they can relate to.

Data science research methods
The availability of large amounts of data also brings a paradigm shift to scientific methods. Instead of relying on samples of the population, big data allows researchers to do their analysis on the entire population. Traditional research is done using statistics to derive theory driven models, while research using big data is more data driven. There is much debate about the validity of these research methods in literature. In this section challenges relating to data science research methods will be discussed.

- Theory and traditional methods(Carter & Sholler, 2016)
- Objectivity and disinterest (Carter & Sholler, 2016)
- Question bias (Rose, 2016)

Carter & Scholler (2016) have summarized discussions that is being done by proponents and critics of data science. In their paper they compare the statements of these two groups with the opinions and practical views of data scientists.
Proponents of data science claim that the availability of data eliminates the need for theory and traditional methods. It is not required to understand the underlying mechanisms and have models of the world, since specific questions can be answered by looking at the data. Critics argue that by not using traditional methods together with the new methods the analysis is less effective. The data scientist interviewed argued that traditional methods are mainly useful in early stages of the analysis. Involving domain experts to spread knowledge and provide feedback to the data scientists should mitigate this challenge.
Another controversial aspect of data science Carter & Scholler recognise is related to objectivity and disinterest. Critics argue that data science methods are claimed to be objective and have disinterest, while they are actually subjective and require interpretation. People tend to perceive information based on computation as facts. When data scientists are able to influence decision making systems that are perceived as being objective, that imposes a risk of manipulation. The data scientists interviewed are well aware of this, and their pragmatic view is that the data is objective, and the interpretation of the analysis is subjective. Keeping the stakeholders informed about the objectivity and the limits of the analysis should mitigate the risk of manipulation.

Rose (2016) has a more practical view on the same problem. He argues that data science is all about asking questions, and that a challenge is avoiding question bias. A data science team needs the right culture and mindset to ensure that the right questions can be asked. He has identified four common causes for question bias.

- Self protection
- Not enough time
- Not enough experience
- Corporate culture discourages questioning

Most people have a tendency to protect themselves. Asking a question to a colleague who has presented an answer makes you vulnerable. But asking simple questions can actually help the colleague in improving their own understanding, which will help the entire team. Asking questions generates more work for the team and also leads to more questions. When a team is under time pressure they can stop asking questions to manage their own workload. The organisation might have a lot of requests in regard to data availability and data quality. Not managing these requests may lead to a situation that the data science team is over focussed on delivering to the organisation and doesn't have time to ask questions. Asking questions takes experience and a different mindset than some people have learned during their career. Team members from software development, engineering or project management have spent their entire career to become the person who has the answers. They tend to ask questions that steer the discussion in a certain direction, which is not really useful. Having a team that have diverse background is important in avoiding this.
Sometimes the question bias originates from when the corporate culture discourages asking questions. A lot of companies are focussed on actions to solve problems. This culture discourages asking questions to understand the underlying causes of the problem. The nature of data science is about asking questions to understand a company's problems. A

culture which discourages this can cause question bias which makes the data science team less effective.

In this section the challenges regarding data science research methods have been summarised. Carter & Sholler have made an overview of the discussions on data science that are being held in the scientific community and society. The pragmatic views of data scientists relating to these discussions provide a more nuanced view that compromises the viewpoints of the critics and the proponents. The challenges defined in this section are far less practical than other challenges in this chapter, but they give a good overview of the risks of data science that is being experienced by society. They are related to the fears and prejudice that may arise when companies change their business using data science. The advantages of data science have to be shown to experts who have been working with their own methods for years. And at the same time they must be convinced that data science and big data will not make their jobs obsolete.
Rose has practical tips to deal with question bias, a key challenge that arises from the nature of data science research methods.

Privacy & ethics
Big data and data science are topics that have some controversy in society. Critics raise their concern regarding certain aspects of big data and data science like data privacy and potential abuse of power by governments or big technology companies.

- Ethics framework (Drew, 2016)
- Privileged access to data (Carter & Sholler, 2016)

Drew (2016) has created an ethics framework for data science in governments. They argue that this ethical framework is required to ensure that people are willing to share their data with government agencies. It should give the public some confidence that the government will handle their data with care, and that they will not do things that can work adversely for some citizens. From this framework challenges for data science specific to government organisations can be derived.

- Minimise intrusion
- Being transparent
- Involving citizens

Most governments have rules to restrict and regulate the use of citizen’s data. This is different from most non government organisations who can use most of their data for data science without limit. Using citizen data for analysis when its not necessary is perceived as an intrusion of privacy. This is a challenge for government organisation who want to do data science.
Governments are supposed to uphold the interests of all citizens. This means that the government should be transparent about what it does and why. That includes being transparent about the data science projects a government organisation does. At the same time it's also undesirable to publish sensitive data.
When citizens see the advantages and results of data science projects they will look more favorable on the use of their data. Involving citizens and showing them results of data science projects is important in government.

Carter & Sholler (2016) identify an ethical challenge in data science. Proponents of data science argue that availability of data will increase democracy and will allow everyone to do research. Critics argue that data is often controlled by big companies which increases the power of these companies over their customers. Companies know that having data grants them power, but some argue that companies should make their data available to the public so that society can benefit.

In this section some of the challenges relating to privacy and ethics have been described. This is especially important for government organisations that want to apply data science, since society expects them to respect privacy and have a higher ethical standard.
The data that is being gathered is less open than people had hoped it would be. “Data is the new oil” is a saying that illustrates the way commercial companies handle their data. They keep it to themselves and use it to gain competitive advantage.
Governments are taking the lead in the open data initiatives. While other companies are more protective of their data government want to publish data others might find useful.

3.5) Results of literature study

The goal of this chapter is to give an answer to the research question Q1:

> ***What are the challenges and opportunities of doing data science projects according to literature?***

The result of the literature study is the concept matrix listed at the start of section 3.4. The following concepts were discovered in the literature:
Opportunities:

- Measure things in greater detail
- Data driven decisionmaking

Challenges:

- Starting data science projects
- Data science team dynamics
- Company mindset
- Data science research methods
- Privacy & ethics

The concepts are high level topics that are composed of multiple challenges mentioned in literature. This list of concepts together with the explanation of each concept has been used as input for the investigation framework.

4) Data science in practice

In preparation of this thesis, a preliminary study at Xomnia was done. The goal was to uncover the nature of the problems that stakeholders of Xomnia experience in the data science domain. The focus of this study is somewhat different from this thesis, but there are relevant findings that support the research framework. The context of the study will be explained in this paragraph, and relevant information will be presented and related to the case study.
One of the problems that Xomnia faced is that a lot of projects don't leave the pilot phase, and deploying them to production is difficult. The potential causes of this difficulty are studied by interviewing multiple experienced people at Xomnia. They talked about their experiences in moving projects to the deployment phase and other problems that might be related.
One of the insights was that the deployment issues might not all have technical causes. In order to provide a context, the possible causes were divided into four quadrants based on two dimensions: issues could be technical or organisational, and they are based in Xomnia or at the customer. The potential causes of the inability to deploy models are presented in table 2.

	Customer	Xomnia
Org.	- No support from employees - No support from IT - No support from high management - Dependant on 3rd parties - No overview of data	- Not enough organisation awareness in Data Scientists - Lack of organisational change vision
Tech	- No big data lab present - No data engineering knowledge - Data not accessible by API's	- Deployment of models not yet mastered

Table 2: Results of the preliminary study. Potential causes of the inability to deploy models.

The potential causes that were found in the preliminary study might be relevant for the case study. The first big difference between the two studies is the viewpoint: the preliminary study is about how Xomnia does consultancy, this thesis is about how an organisation does data science. King's Landing has many more data science projects in which Xomnia is not involved. Therefore the Xomnia dimension of the table above is not so relevant. King's Landing is a large government organisation which tend to be bureaucratic. Support is needed on multiple levels to complete projects. 'No support from employees/IT/High management' is therefore an interesting aspect to investigate. The municipality uses domain specific SaaS software for the youth care operations making them dependant on this party to provide data. The municipality already has a big data lab to do their experiments. The other causes are less interesting or relevant to the case.
The most interesting aspects of the preliminary study are:

- **Difficulty deploying models to production**
- **Lack of support from organisation**

These two aspects are the answer to research question Q2:

What are challenges of applying data science in practice?

The aspects are relevant to the case study in King's Landing and can be used in the next chapter to ground the investigation framework in practice.

5) Investigation framework

The investigation framework will provide the design of the case study. The literature study and the research topics will be used to form the base of the investigation framework. This will ground this research both in literature and in practice.
The aim of the case study is to learn whether King's Landing is likely to experience the challenges found in literature and to report any challenges that were not found in literature.

5.1) Based on literature study

Not all challenges and opportunities found in literature might be relevant. In this section, the relevance of these challenges and opportunities will be described. This relevance will be used to determine the interview questions and as a guideline to prioritise follow up questions.

Opportunities

- **Measure things in greater detail**

Measuring things in greater detail is not something that is a high priority for a municipality. Because of the wide range of responsibilities, the processes are highly diversified. This means the added value of gaining deep insight in a specific process is relatively low.

- **Data driven decisionmaking**

Data driven decisionmaking is very relevant for King's Landing. Employees often make decisions based on gut feeling instead of information. The municipality has a high level goal of improving the data-driven decisionmaking in their organisation.

Challenges

- **Challenges in starting data science projects**

King's Landing has asked Xomnia to help them start a project. The goal of this project was also to learn how to start new data science projects. This makes the challenges in starting data science projects highly relevant.

- **Data science team dynamics**

Data science team dynamics are more relevant when the team is a bit bigger. At this time, only Xomnia consultants are working on the project as a proof of concept, while King's Landing employees are only involved for a few hours a week. This challenge would be more relevant when King's Landing is creating their own full time data science teams.

- **Company mindset**

Government organisations are quite bureaucratic and slow to change. There are also a lot of employees have been working within the municipality for dozens of years or more. These employees have developed their own way of working which is based on their experience. This might clash with the new way of working that is enabled by data science. Therefore the company mindset a useful challenge to explore.

- **Data science research methods**

Classical statisticians are being trained to learn data science techniques which means they have to learn a new paradigm. Challenges based on research methods are somewhat interesting and can be pursued if there is an expressed interest in the challenge during the interview.

- **Privacy & ethics**

Privacy and ethics are very important to King's Landing, since they are a government organisation handling sensitive data about civilians. The privacy & ethics challenges in data science are new for all government organisations, and there is no clear policy on how they should be handled.This makes these challenges very relevant.

The relevance of each of the opportunities and challenges found in literature has been explained in this section. Data science team dynamics will be totally omitted from the interviews. The rest will be used to create the interview design.

5.2) Based on research topics

As seen in the previous chapter, interesting challenges based on the research topics are:

- Deploying models to production
- Support from organisation

Deployment of models is out of scope for the case study project. Research topics has shown that there are a lot of challenges to it. Hearing the ideas of the stakeholders on deployment will illustrate the goals they hope data science to achieve.
Getting large, traditional organisations to adapt data science is quite challenging. Support from the organisation is required to ensure it succeeds.

5.3) Project methods

During the interview it is important to talk the same language as the interviewee. The project methods have certain steps, events and milestones that help structure the project. These project milestones and events can be used to get more detailed information about the projects.
CRISP-DM (Cross Industry Standard Process for Data Mining) is the data science methodology that is being used in the project. The steps of CRISP-DM each describe different activities within a data science project. Most of the practical problems or issues arise in a specific step. To get more details during the interviews, the challenges experienced in each step will be discussed. This forces the interviewees to think in more detail about their past experiences.

5.4) Interview structure

The interview structure is designed based on the literature and the preliminary problem analysis. The design of the interview is described in table 3. For the specific questions that were asked during the interview, refer to Appendix A. In the rest of this section, each phase will be discussed.

Phase	Subject	Base in literature	Base in Research topics
Introduction			
Open discussion	Opportunities		
	Chances		
Challenges in current project	Business understanding	CRISP-DM	Support in organisation
	Data understanding	CRISP-DM	
	Data preparation	CRISP-DM	
	Modeling	CRISP-DM	
	Evaluation	CRISP-DM	
	Deployment	CRISP-DM	Deploying models to production
Topics	Data driven decisionmaking	Literature study	
	starting data science projects	Literature study	
	Company mindset	Literature study	
	Data science research methods	Literature study	
	Privacy & ethics	Literature study	

Table 3: Interview design.

Introduction

The start of the interview is about the background of the interviewee. This will include the position in the organisation, background and what knowledge is contributed to the project. It will be useful for readers to know the background of people that are interviewed.

Open discussion

The interview will begin with an open discussion about the challenges and opportunities in applying data science in their organisation. The goal of this is to begin the interview with the interviewee unbiased of literature challenges and opportunities. It also helps to judge the priority of the challenges perceived by the interviewee. When one of the topics of the last part is brought up in the discussion, it can already be explored a bit more in depth.

Challenges in current project

The next part of the interview will be focussed on practical challenges in the youth care project. The interviewee will be asked questions for each of the steps in CRISP-DM. The research topics has shown that support in the organisation and deployment of models are relevant challenges in other projects of Xomnia.
Depending on the person interviewed, this phase can be a big or a small part of the total interviews. People closer to the project will have more relevant input than people that are in a high level position in the organisation.

Topics

The list of topics is based on the concepts that result from the literature study. They can be used when an interviewee expresses interest in the topic, or when the interviewer thinks the interviewee has an interesting opinion on the topic. Not every topic will be of interest to every person.

5.5) Results of the investigation framework

The results of Q1 and Q2 have been used to decide what challenges are worth studying in the case study. Based on these challenges an interview design has been made.
The interview design is the answer to research question Q3:

> ***What challenges are expected to be important when doing data science projects in municipalities?***

The interview design will be used to collect case study results for Q4.

6) Results

In this chapter the results of the case study will be presented. The combined interviews are summarised and accompanied by quotes of the interviewees. The transcripts of the interviews will not be published to protect the confidentiality of King's Landing.

6.1) Interviewee backgrounds

To get a complete view on the case, people with different backgrounds and roles were interviewed. Different backgrounds and roles will result in different views on the problem. As a result, the interviews had a slightly different focus. People higher in the organisation often generalise challenges and experiences over multiple projects across the entire organisation. At the same time, people closer to the project can give more concrete examples about the youth care projects. By combining the input, this research aims to give a complete picture of the challenges experienced in King's Landing. The backgrounds of the interviewees are displayed in table 4. Person C and D were interviewed simultaneously and the rest separately.

Person	Organisation	Position	Role
A	JGZ	Manager of two youthcare locations, Product owner	Product owner in JGZ project
B	Municipality	Project manager data-driven decisionmaking	Aligning project goals, focus on results
C	Municipality	Strategic advisor at dept of information management	Overseeing all projects, guarding coherence
D	Municipality	Program manager data-driven decisionmaking	Responsible for all data-driven decisionmaking projects
E	Xomnia	Data scientist	Data scientist in JGZ project

Table 4: Backgrounds of interviewees

6.2) Project context

The case study will be about a Dutch municipality with about 300,000 inhabitants. In this paper the municipality will be called King's Landing. King's Landing is doing several projects regarding data science and advanced analytics. In the case study one project will be studied, which is about youth care.

In the Netherlands, the government has programs in place to monitor the health and development of children aged from 0 to 18. These programmes are distributed to the municipalities and the organisations responsible are the *Jeugdgezondheidszorg* (JGZ; "youth healthcare" in English). The JGZ is responsible for planning and executing the monitoring of the youth, which is takes place in around 17 contact moments. In these contact

moments different healthcare professionals handle tests and measurements relevant to the life stage of the child.
There are questions within the JGZ about how they can improve and possibly diversify their services. Now they offer one route through their programs which every child takes, while some children might need more attention than others. In the rest of society this is going on for some time, and by making use of technology they could diversify their services, or target specific groups of citizens which need a certain type of care. As a part of these innovations the JGZ has initiated a project to explore the possible benefits of using data science. They have epidemiologists who study the health of the population and detect trends among the population. They often base their research on surveys and data collected specifically for their research. Another goal of the project is to educate these epidemiologists in modern data science technologies.
The JGZ has hired Xomnia to help them with the project and the training of their employees. Xomnia is a data science consultancy company with around 40 employees based in Amsterdam. Xomnia has supplied data scientists to help with the project and assess the training required by the employees of the JGZ. They also have a portfolio of trainings that could be offered to the JGZ to improve technical skills and knowledge of employees.
The JGZ has acquired a virtual datalab from a hosting company. This datalab is a secure Hadoop environment that has all tools necessary to conduct data science projects. The JGZ will give Xomnia access to this datalab, so all sensitive data stays within their own secure environment. This datalab is new for the JGZ and Xomnia will train the employees of JGZ in using the datalab.
An important stakeholder in the project is the security officer (DISO). He is responsible for the security and privacy aspects of all projects within the municipality. Therefore he has to approve many different aspects of the project: which data is used, who has access to it. The team has to write plans for many of the different steps which will then be approved by the DISO. When the team needs more data they have to write a motivation on why they need the data.

6.3) Project challenges

During the interviews the practical challenges in the youth care project were discussed. The amount of useful input varies per person, as people closer to the execution of the project know more about practical aspects.

Business understanding

The business understanding was where most of the difficulties were experienced in the JGZ project. The first step of the project was a session with the product owner, one domain expert and the data scientists. The data scientists also visited the JGZ location where all the youth care takes place. Later in the project, demo sessions with multiple domain experts were organised. During all these sessions, it was difficult to define a clear research goal suitable for data science. One of the possible causes is the lack of knowledge of the stakeholders.

> *Data scientist: 'Wij zijn als data scientists afhankelijk van de mensen op de werkvloer voor het krijgen van input en feedback. In dit project zijn de specialisten er bijgehaald zonder dat ze voorkennis hadden over wat data science zou kunnen bieden. We hebben ze dit tijdens de demo sessies moeten leren.'*
>
> *Data scientist: 'As data scientists, we depend on others to get feedback and input. In this project, the medical specialists had no prior knowledge about the opportunities data science brings. We had to teach them these things during the demo sessions.'*

The result was that creating useful data science research questions was difficult, as they tended to be too general. healthcare professionals are interested in the quality of healthcare, but have no ideas about how this could be operationalised.

> Project manager: *'Alles valt of staat met het hebben van een goede onderzoeksvraag, als je begint met wat aanrommelen met data, dan wordt het nooit meer dan dat.'*
>
> Project manager: *'Having a good research question is important. When you are just messing around with data, that is all it will ever be.'*

In the end, the data scientists managed to define research questions which could help the JGZ, but it took a lot of effort and there is definitely room for improvement.

Data understanding

Together with an epidemiologist of the JGZ, the data scientists formed a list of useful variables to include in the analysis. The data scientists could communicate with two people from IT operations to clarify any questions they had. In cases where IT operations couldn't help, the data scientists could contact the domain experts instead.

One issue was that certain attributes in the dataset were missing many values. The data scientists deducted it could mean two things: the healthcare professional either did not check it, or found no problems. The source system did not clearly distinguish between these two cases.

> Product owner: '*Een veld over de lever en milt werd erg weinig ingevuld door professionals. Tijdens een gesprek met artsen en verpleegkundige bleek dat deze velden een afgeleide waren van een check op de buik. Als er geen problemen zijn met de buik dan wordt het veld over de lever en milt niet ingevuld. Dit betekent dat het goed is, maar aan alleen de data is dit niet te zien.'*
>
> Product owner: '*A field containing a check on the liver and spleen had a lot of missing values. During a conversation with healthcare professionals we found out that this check was part of a check on the stomach. When it is empty there are no problems, but this is not apparent from the data'*

Data preparation

The quality of the data was really low, which caused the data preparation to take a lot of time. Each child has multiple contact moments which are registered in the system. Data science models require the data to be in a row format with a constant number of columns. Therefore the multiple contact moments had to be aggregated into several constant variables. This difference in the dimensions made the data preparation complex and time consuming.
Another complication was that a lot of fields are free text fields. A field containing the duration of a pregnancy would naturally be a number. In the data it turned to vary from minutes to weeks to months, so data scientists had to do a lot of transformations to convert this free text field to a standardized unit of measurement.

Modeling

When designing a predictive model, a data scientist typically has to select one variable that must be predicted. This variable is often called the “label”. During the JGZ project, defining these labels was difficult, as the business could not define them. The business wanted to predict healthcare quality but that is abstract and quite complex. In the end the data scientists and the business defined three possible labels:

- Special education: is the child receiving special education?
- No show ratio: ratio of people not showing up to appointments
- Average number of extra appointments: Extra appointments are scheduled when something is ‘wrong’.

Different models were trained with the labels above. Data scientists selected features and tweaked the parameters of the models to improve their performance.

Evaluation

There are two forms of evaluation in data science: statistically measuring the model performance and letting stakeholders evaluate the usefulness of the model.
The special education label was not predictable at all, so it was discarded. The no show ratio had mediocre performance compared to the other models, and the average number of extra appointments was fairly good but not great.
The data scientists held demo sessions with stakeholders to evaluate their usefulness. In data science, the performance requirements always vary per domain and application. The task of the data scientist is to find out what these requirements are in a particular situation.

> Data scientist: *‘Het kan zijn dat een specialist een model gebruikt om zijn onderbuikgevoel te controleren, dan stel je minder hoge eisen.’*
>
> *Data scientist: ‘When a model is only used to verify the gut feeling of a medical specialist, demands for model performance can be lower.’*

Deployment

Deployment of models was not in scope during the project. The interviewees still had ideas on how to deploy the models. There are two main areas of application: healthcare operations and policymakers. These two areas require different kinds of deployments.
When a child has an appointment, the professional consumes information from an information system. A predictive model could control a flag or notification in the information system, so the professional can consume the insights in the place that is natural to them. Implementing this will be quite complex since the cooperation of the information system supplier is required.

Product owner: '*Een model deployen voor de professionals op de werkvloer zal via het dossier systeem moeten. Hiervoor zijn we erg afhankelijk van de leverancier.'*

Product owner: *'Deploying a model for healthcare professionals should go through the system of records. For this we are highly dependant on our supplier'*

The other option is providing insights and predictive models to policymakers. They are more interested in long term effects and trends. Displaying the results of a model in a dashboard, graph or report could help some employees make better decisions. This method of deployment is also less complicated than the previous method.

6.4) Perceived opportunities

The interviews revealed various opportunities as they are perceived by the interviewees. They can be grouped into several categories:

- Data driven decisionmaking
- Improve services
- Data driven Policymaking

Data driven decisionmaking

Within the municipality of King's Landing there are programs to improve data-driven decisionmaking. In the interviews it became clear that data science has a prominent role in improving the data-driven decisionmaking. There were different perceived opportunities that can be categorised as data-driven decisionmaking.

Internal

When everyone in the municipality has access to relevant data and information they can make better decisions in their daily work.

> *Strategic advisor: 'Door data science toe te passen kunnen we ook onze interne bedrijfsvoering een stuk efficiënter doen. Wanneer de medewerkers in alle lagen realiseren wat de waarde van data is, is mijn programma geslaagd. Ze kunnen dan zelf bedenken hoe ze met data hun werk beter, sneller of mooier kunnen invullen.'*

Strategic advisor: 'By applying data science we can improve our internal efficiency. When employees across the organisation realise what data can offer, my work is successful. They will be able to decide how data can improve their work.'

External

Every municipality has a coordinating role within a range of different supply chains. These days they are seen as an equal partner instead of the bureaucratic organisation that decides everything. When a municipality can provide the supply chain partners with relevant data they can make the whole supply chain more efficient.

> Program manager: *'We hebben binnen de gemeente erg veel data en ik denk dat wij met die data in de keten ook waarde kunnen toevoegen.'*

> Program manager: *'We have a wide variety of data within the municipality, and I think we can use that data to add value in the supply chain'*

Increase data quality

During the project at the JGZ they found out that the data quality is not really good. The project manager pointed out that data science can also be used to increase the data quality within the organisation.

> Project manager: *'Door naar je proces te kijken met data science ontdek je plekken in de operatie waar slecht geregistreerd wordt.'*

> Project manager: *'By studying business processes with data science we can uncover where people aren't registering properly'*

Improve services

Several interviewees mentioned improving the services the municipality offers. In general they saw an opportunity to improve the quality, and also an opportunity to make their internal processes more efficient.

Quality

Citizens of King's Landing make use of the services that the municipality provides. Improving the quality of these services is one of the opportunities that were mentioned during the interviews. One of the ways to do this would be to customise the services to match the needs of individuals.

Efficiency

There is a lot of data in transactional systems of the municipality. This data can be analyzed to find processes that are suboptimal or inconsistent.

Data driven policymaking

A lot of the activities within the municipality have to do with policies. Policies are the mechanism the municipality uses to interact with the environment. During the interviews there were a lot of opportunities that are related to policies.

Less assumptions when creating policies

When employees of the municipality create policies they have to make a lot of assumptions. Often the policy is about an abstract, hard to measure concept like ‘loneliness’. When they have more data and information they can verify their assumptions by comparing them to actual data about related concepts. By doing this the result of their work should be more accurate.

Insights in citizens priority

Employees and politicians within the municipality often have certain ideas about the priority that issues have. By applying data science it is possible to check the priority citizens give to those issues. Counting the complaints of citizens and grouping them by topic are simple solutions that are possible with data science.

> Product owner: *‘We hebben ons ooit gefocust op een bepaald probleem, maar na een tijdje bleek dat burgers dat helemaal niet ervaarde als probleem.’*
>
> Product owner: *‘We focussed on a certain issue, but after a while we found out that our citizens don't perceive that issue at all.’*

Feedback on new policies

Currently when the municipality issues a new policy they have to wait for a year to receive feedback on it. This feedback is based on surveys and reports of people executing the policy. During the interviews several people thought that data science could be used to measure the effect of a certain policy in real time. This will allow the municipality to adapt ineffective policies on short term instead of having to wait a year.

Better execution of policies

The policies that are defined by the municipality also have to be executed by the municipality. This is often done in collaboration with partners in a supply chain, but just like other processes data science can be used to do it more effective.

> Strategic advisor: ‘*Doordat je meer informatie en een gelijkwaardige relatie met ketenpartners hebt kun je gericht naar problemen in de keten kijken.*’
>
> Strategic advisor: *‘Having more information and an equal relation with partners allows us to gain insight of issues in the supply chain.’*

From the interviews it appears there are several opportunities applying data science could give King’s Landing. Some are specific for municipalities and others are also applicable to other organisations.

6.5) Perceived challenges

During the interviews there were several challenges in applying data science that are relevant to King's Landing. In this section each challenge will be summarised based on the interviews.

Data quality & quantity

During the experiment at JGZ in turned out that the data was not suitable for data science applications. There are a lot of missing values and free text fields. For a healthcare professional entering free text is the fastest way to register something, but for data science more structure and consistency is better. Duration of pregnancy is a field which could be highly structured as the number of days. But it was in a free text field which lead to health care professionals using a different units, varying from days to months.
Missing values and free text field make the data preprocessing more time consuming and decreases the predictive strength of models.
There was also much less data than everyone expected. In the experiment data spanning 6 years was used to train and test models. The result was that there is no data which spans the entire youth (birth to 18th birthday) of a child.

> Product owner: *'We hebben 6 jaar aan data gebruikt in het huidige experiment, en dat lijkt ontzettend veel. Als je ziet wat de voorspellende waarde is die daar uit volgt dan valt dat erg tegen.'*

> Product owner: *'We used 6 years of data in the current experiment, which seems like a lot. But the predictive power of the models based on the data was quite low.'*

Data Access & Tenders

One of the challenges for data science in general is that all data is spread out throughout the organisation. Every department has their own system of records which creates isolated silos. Making the data available and combining it with data from other parts of the organisation could be beneficial.
When looking at long term analysis another problem is that the municipality has to tender every couple of years. When the source systems change periodically it is difficult to build a long term representative dataset.

> Product owner: *'Het kinddossier moet worden aanbesteed. Als de volgende aanbesteding door een andere partij dan de huidige wordt gewonnen dan moeten we nog zien of alle informatie goed over te zetten is.'*

> Product owner: *'The system of records is subject to a tender. If the next tender is won by a different supplier I hope we can keep all historical data'*

Privacy

Privacy and data science is a controversial topic within the municipality. Everyone talked about this challenge during the interviews. The baseline is that the municipality values the privacy of its citizens a great deal. At the same time they see the opportunities that data science offers. Unlike many commercial companies they want to be conservative when it comes to privacy.
At this time there is much uncertainty when it comes to what is allowed according to the new privacy laws. This is highly disruptive to data science projects because this uncertainty results in a paralyzing fear. And because of this fear the municipality does not learn how to interpret the privacy laws which results in continued uncertainty.

> Strategic advisor: *'Er is nog erg veel onzekerheid over wat wel en niet mag binnen de kaders van de privacywet, dat zorgt voor veel onrust. Om de zoveel tijd komt er binnen een project iemand van een congres met een whitepaper van 40 pagina's over privacy. Dan gaan mensen die er geen verstand van hebben informatie interpreteren waarvan je niet weet of het juist is. Dan krijg je binnen je data science projecten 'the blind leading the blind'.'*

> Strategic advisor: *'There is much uncertainty on what is allowed within the confines of the privacy law. This causes a lot of unrest. Every now and then someone in a project goes to a congress, and receives a 40 page white paper about privacy. People without knowledge interpret information that might be incorrect and the result is the blind leading the blind.'*

Ethics

During the interviews many people elaborated the ethical nature of applying data science. The main question is about whether the municipality should do data science at all. Just because technology allows it does not mean it should be done. During the interviews there was no clear answer for this question.
In case of the JGZ project this is even more sensitive. One of the core goals of the JGZ is to be easily accessible to all citizens. They want people to feel like they can trust the JGZ and ask them for help when they have problems. When the JGZ would use data science to predict the needs of children, the model could unknowingly discriminate against a certain part of the population. Even if it could potentially help these children the question remains whether it desirable to profile people using data.
In the project there was an ethical committee that was tasked with making decisions on how to act based on possible insights. By doing this the JGZ hopes to do data science while still safekeeping the rights and interests of the citizens.

Changing the way of working

The strategic advisor and strategic manager are also involved in other projects and they had some interesting observations regarding the interaction between data scientists and experienced, older employees of King's Landing. They both believe data science can be

used to radically change the way of working within the municipality. Using data to improve policymaking is one of their main objectives.
But they see that the collaboration between data scientist and policymakers can be challenging. The policymaker is used to thinking about a problem and writing a piece of text. The data scientist wants to look at the data to better understand the problem. The policymakers can get nervous from this ad-hoc way of working. Facilitating the collaboration between these two groups is one of the challenges the municipality has to overcome to do more data-driven policymaking.

> Program advisor: *'Van de snelle ad hoc manier waarop data scientists werken worden beleidsmedewerkers zenuwachtig. Dan grijpen ze bijvoorbeeld naar een argument als privacy om te zeggen dat iets niet mag.'*
>
> Program advisor: *'The ad hoc way of working can make a policymaker nervous. Sometimes they resort to using privacy as an argument to block progress'*

Data science research question

One of the most important challenges in the JGZ project was how difficult it was to find a good research question. The possible causes are also described in chapter 6.3 in the business understanding paragraph. Often the questions the business wanted to know where too abstract and it was very difficult to get actionable useful research questions. At other times the research questions didn't help to solve a specific problem.

> Program advisor: *'Voordat we met het JGZ project starten zijn we erg lang bezig geweest met het helder krijgen van de vraag, en dat was erg lastig. En toen we het project uiteindelijk starten is het een hele andere vraag geworden. Dat laat maar zien hoe moeilijk het helder krijgen van de vraag is. Uiteindelijk hebben we met Xomnia wel stappen gezet.'*
>
> Program advisor: *'Before we started the JGZ project we spent much time on clarifying the main research question. When we finally started the project the question shifted to a totally different one. This shows how hard it is to pinpoint the right question. Xomnia did help us with this process.'*

Acting on insights

The project manager noted that when insights are acquired the organisation still has to act upon them. He observed during a session that the business was well aware of some issue that the data scientists had found in the data. If the issue was known, why hadn't they acted to solve it? If the organisation does not act upon the insights generated by data science it is a waste of money.

6.6) Other Challenges

In the phase 3 of the interview the challenges found in literature were presented, and the interviewees were asked if they recognised the challenges. Some of the topics were already discussed during other phases of the interviews. Most of the other topics did not provoke an interesting response. This was probably because the challenges and topics did not (yet) apply to the case. The interview was designed so that these topics could be used to structure the conversation and get additional information at the end of the interview. In practice at the end of the interview most important topics were already discussed.

6.7) Privacy

Privacy was an important topic during the case study and the approach in King's Landing deserves some extra attention. There is a lot of uncertainty about what is allowed by the new dutch privacy laws which often cripples innovation. Meanwhile they do want to run some experiments to let the organisation get familiar with data science. During these experiments the privacy of citizens and security must be guarded.
During the experiments there is a range of security measures and processes in place. A security & privacy officer is appointed to approve all data science projects and even specific variables that will be used. This person will make sure that the privacy is guarded well and that there is a clear goal to justify the use of data.
To ensure that sensitive data is not lost or stolen King's Landing is using a highly secure data lab. In this data lab the data scientists can run their experiments without having to store data on their local machines. This reduces the risk of data being lost or stolen.
These measures ensure that the current experiments are done in a secure manner. By running these experiment King's Landing also learns more about the practical aspects of privacy.

These practical lessons are quite useful, but at the same time King's Landing hopes to gain insights in the new privacy laws from the legal perspective. For the next step they will work together with a law firm to study the new laws to define and clarify how they should be interpreted. In the future they hope to have a framework which defines the gray area between what is allowed and what is not. Decisionmakers can use this framework to make conscious decisions on when the privacy risks are worth the possible benefits.
This should reduce the fear and uncertainty around data science in the organization. The people involved in projects can use the framework to see if something is allowed. When it falls in the grey area someone can make a decision.
Privacy issues are a big challenge for King's Landing when doing data science. From the perspective of the citizen it is good that they are conservative. They have taken measures to mitigate the risks but it is still an obstacle which slows down data science projects.

6.8) Summary of results

The results described in this chapter give a clear picture of the challenges and opportunities that the interviewees perceive. Due to the different roles they each had in the project their

input gives a complete image from multiple standpoints. The results described in this chapter answer research question Q4:

How does King's Landing address these challenges (found in Q3)?

The data scientist and product owner managed to give a clear picture on what went on in the project in each stage of CRISP-DM. **Business understanding** was the most difficult phase because it was hard to define good research questions. The **low data quality** was another main challenge in this project, which made data preparation difficult, and model performance bad.

The project manager, strategic advisor and strategic manager gave insights in the challenges and opportunities perceived for the entire municipality. data-driven **decisionmaking**, **improved services** and data-driven **policymaking** were identified as the most important opportunities. The most important challenge is **privacy** because unclarity in the law creates a paralyzing fear of data science within the municipality.

7) Discussion

In this chapter the organisation maturity models will be related to the results of the case study. The generalizability of the results will also be discussed.

7.1 Organisation maturity

In the case study project King's Landing aims to learn how they can best apply data science. The gartner maturity model, described in the introduction, differentiates between hindsight, insight and foresight. Hindsight means employees are using historical data to understand the past. Having Insight means employees can combine data to understand the present. Foresight refers to being able to predict what will happen.
When taking a pragmatic approach, reaching a higher maturity level requires two things:

- Better data infrastructure (Quantity, quality and variety)
- Better data-analytic capabilities.

To illustrate these requirements consider the following comparison: Training a predictive model often requires multiple data sources that have all been cleaned and preprocessed. Creating a simple revenue chart requires data from a single transaction system. Most employees in an organisation can comprehend and use simple metrics like revenue. A more complex analysis like a prediction is more abstract and harder to comprehend. When the analysis gets more complex, the users also need better data-analytic capabilities.
In this section the two requirements will be elaborated using existing maturity models: data management maturity models and analytics maturity models. The challenges encountered in the case study will be related to these maturity models.

Data management maturity

Organisations have seen the importance of data management for efficient operations. Data analytics can benefit from having mature data management. The CMMI institute has developed a data management maturity (DMM) model to support improving data management capabilities in organisations. (Mecca, 2013) The main concepts in this model are displayed in figure 6. In this section this model will be used to discuss the effects of data management maturity on the challenges in applying data science.

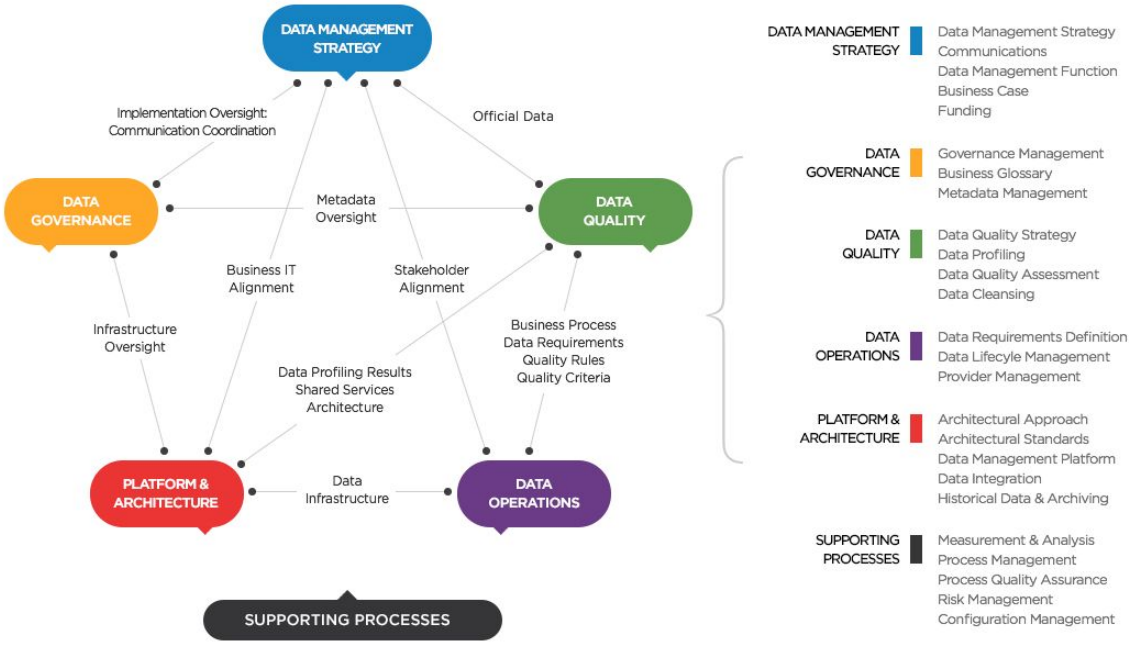

Figure 6: Data Management maturity by CMMI institute. (Mecca, 2013)

Mapping the DMM was not the goal of the interviews, but there is still some relevant information regarding the DMM of King's Landing. Several of the aspects of DMM are relevant in the case study at King's Landing: Data management strategy, data governance, data quality and Platform & Architecture.

The program manager and the strategic advisors are busy with forming the ***data management strategy***. The goal is to increase the data-driven decisionmaking and to become an information partner to their supply chain partners. They are extending the strategy to include data science by using the lessons learned from the data science projects.

Data governance is on a much lower level. The different departments all use different IT systems and there is no central metadata repository. This makes it difficult to find data around a specific topic. Having central metadata management helps data scientists find the data they need. Having data governance in place should also reduce the privacy risks in applying data science. Mature data governance includes having well founded data access levels and roles which define what data users are allowed to see. It was difficult to configure the big data and data science tools to use these advanced access controls.

Data quality was a big challenge in the case study. When the data quality is monitored the data scientists know on beforehand what the quality of a certain dataset is. When data quality is improved on the source and infrastructure level the results of data science models will benefit from that. It is also much more efficient to improve data quality in a central place compared to letting data scientists do that on a project basis.

A clearly defined ***architecture*** and well implemented ***platform*** is required to realise all the other aspects of a data management strategy. In King's landing there is no centralised data warehouse where all relevant data is ready to use for analytics. When doing data science it is also preferable to have a data science platform to do experiments in, alongside the regular data infrastructure. During the case study it became clear that a data science platform that is mature and flexible to the demands of the municipality was not yet available.

There was little time during the case study to discuss the data maturity of King's Landing. Therefore an exact statement about a maturity level is not useful and also not relevant for this research. What can be concluded is that many of the challenges that were encountered during the case study project would not have happened with a higher DMM.
Mature data management can also help provide safeguards related to protecting the privacy of citizens. When there are existing clear rules, processes and infrastructure related to data management, enforcing privacy related governance models is as easy as adding a few more rules and processes.

Analytics maturity

The analytics maturity of an organisation can be defined as the capability to use data and information for improving decision making. In the case study there were a number of challenges that can be explained by King's Landing's analytics maturity.
The focus of the gartner maturity model mentioned in the introduction was the type of questions that are asked. There are several other maturity models that have their focus on the organisation. Business intelligence has been around for a few decades and successful usage of BI in an organisation is an indication of how open business users are to data-driven decisionmaking. Gartner published a maturity model on business intelligence in 2009 which is not related to big data but still relevant. (Burton, 2009) TDWI has published a maturity model that is more focussed on the adaptation of big data within an organisation. (Eckerson, 2007) Although there is a slight difference in these models, they have different levels and the implication for the organisation is very similar. The different maturity levels are summarised in the table below:

Level:	Gartner	TDWI	Description
1	Unaware	Nascent	Ad hoc data analysis, not reusable and no DWH infrastructure
2	Tactical	Pre-adoption	Plans for DWH infrastructure started, organisation becomes aware of analytics.
3	Focused	Early Adoption	Several successful projects, awareness grows
4	Strategic	Corporate Adoption	Analytics change how companies do business. Analytics is a core component of the global strategy
5	Pervasive	Mature / Visionary	The use of analytics has enabled the organisation to gain significant competitive advantage. The organisation becomes a leader within its sector.

Table 5: Analytics maturity levels by Gartner and TDWI.

During the interviews there was little focus on the analytics maturity, but analytics maturity can be used to explain some of the challenges experienced in the case study project. Based on the information acquired during the interviews it would appear that King's landing is somewhere between level 1 and 2. According to the program manager one of the long term

goals of King’s Landing is to use data science to change the way they make policies. That goal would match with maturity level 4 which does not match the current situation at King’s Landing. So the ideas are there but there are several impediments to realising them, as was found out during the case study.
They have no central data warehouse that holds all relevant data in the entire organisation. The application that held relevant data for the project had standalone reporting capabilities. Assuming this is a representative situation one can assume that a lot of other departments in King’s Landing also have the data in silo’s. The data siloes need to be opened up before higher maturity levels can be reached.
During the case study project it became clear that the stakeholders involved to provide domain knowledge are not used to working with data. A characteristic of higher maturity levels (3+) is that data-driven decisionmaking is used in the entire organisation. Defining a data science research question was more difficult than most interviewees had imagined on beforehand. Maybe this can be explained by the lack of analytics understanding of the stakeholders. Several years of experience using historical data for decision making could make a difference in the process of defining useful research questions.

King’s landing has a relatively low analytics maturity. There is no centralised data warehouse, data is in silo’s and large parts of the organisation don't use historical data to base their decisions on. The successful application of data science is probably dependant on having a higher analytics maturity. Creating a data warehouse to centralise data from all data siloes should enable employees to use historical data to base their decisions on. This will also give them the experience in basic data-driven decisionmaking to be involved in more advanced analytics applications like data science.

Maturity models and data science

In this section the results of the case study in King’s Landing have been related to process and analytics maturity models. Maturity models are useful tools to help organisations determine the next steps to take in order to improve. Data science is a relatively new field for which no established maturity models exist. The data science field overlaps with other fields like Analytics and IT, so those maturity models can be used.
The results of the case study suggest that following the progression of maturity models could help the effectiveness of data science projects. More research is required to find out if the existing maturity models are enough to cover the data science field.

7.2) Generalization

One of the goals of this research is to gain insight in how organisations apply data science. During the case study the situation of a specific organisation was studied, and it is interesting to see how the findings can be generalized to other organisation. Findings can be generalizable to a certain type of organization or to a specific domain.

Government organisations

King's Landing is a municipality and other government organisations will probably behave in a similar way. They have no competition so they have less motivation to innovate quickly. At their heads are politicians who fear negative publicity much more than regular executives, making them more conservative.
The privacy issues King's Landing experienced are probably also a challenge for other government organisations. The new privacy laws combined with a new field like data science give uncertainty. Other government organisations operate under the same laws so they will probably face similar challenges.
Government organisations don't have external pressure for innovation that is caused by competition. They serve the interests of all civilians which gives them the opportunity to regard the ethical aspects of data science. A widely present opinion in King's Landing was that when technology allows the use of data science to improve something, that is not an argument to do so. These ethical considerations are probably also happening in most government organisations. This attitude is desirable in government organisations, but it does slow down data science adaptation.
Several issues experienced at King's Landing can be associated with being government organisation. This makes the results of the case study largely generalizable to other government organisations.

Health care domain

The data science project in the case study was about the healthcare domain. The healthcare domain is also conservative when it comes to the privacy of their patients. The laws about privacy are often more strict for medical records. The findings and challenges regarding privacy are therefore generalizable to some healthcare organizations.
During the project it was difficult to come up with useful research questions. Due to the complex nature of healthcare the professionals are used to using abstract measures like 'Quality of care' which are hard to measure. Translating these abstract measures to something the data scientists can make using the data available was difficult. These difficulties are probably also relevant when doing data science in other healthcare organisations.

Large organisations

Some of the findings in the case study are generalizable to large organisations. King's Landing is a large organisation that has several large semi-independent departments. A side effect is that these departments have their own isolated information systems and have little interaction. This results in data 'hidden' in silos across the organisation. Within departments there is also little knowledge about the data that is available in other departments.
In King's Landing siloed data was a large obstacle in making the organisation more data driven. It makes data science difficult because the data scientists have to spend more time finding and extracting the data.

The ultimate goal of data science is to improve decision making by using data. When all the data is available and exposed to the organisation, employees still have to use it in their decision making. This requires change management, which is more complex in large organisations and is required to reap the benefits of data science.
This is well illustrated by one of the concerns of a project manager in King's Landing. He thought it was important to stimulate people to act on the new insights they get. When a department is not acting on existing insights, providing them with more detailed information will not provide benefits to the organisation. A change in employee behaviour is also required.

Future work

The statements made in this research would be more reliable if backed by empirical evidence. An empirical study that follows multiple organisations over a longer period of time would give deeper insights in the application of data science. A clearly defined measurable end goal on business level, combined with information on the approach of each project would give a lot of insights on how to make data science projects succeed.

8) Conclusion

This research aimed to find answers to the following questions:

Q1 What are the challenges and opportunities in doing data science projects, according to literature?
Q2 What are the challenges of applying data science in practice?
Q3 What challenges are expected to be of importance when doing data science projects in municipalities?
Q4 How does King's Landing address these challenges?
Q5 What could we learn from King's Landing's experiences?

The answer to Q1 has been found by doing a systematic literature review. This resulted in the following challenges and opportunities that have shown relevance towards the case study:

Challenges:
- Starting data science projects
- Data science team dynamics
- Company mindset
- Data science research methods
- Privacy & ethics

Opportunities:
- Measure things in greater detail
- Data driven decisionmaking

These challenges and opportunities have been described in chapter 3. The answer to Q2 are challenges based on an earlier study at Xomnia, which is the data science consulting company involved in the case study. The following challenges found in the earlier study are relevant to the case study and are described in further detail in chapter 4:
- Difficulty deploying models to production
- Lack of support from organisation

The investigation framework discussed in chapter 5 answers Q3 and is the foundation for the case study design. The insights from the literature review (Q1) and observations from the practical study (Q2) are combined in order to to provide increased focus for the case study. During the interviews, an open discussion was held about the challenges and opportunities experienced when applying a data science. That part of the interview is broad and meant to be about the entire organisation. Another part is about the challenges in the current project which provided practical and specific insights.
During the case study the challenges King's Landing faces are studied. In chapter 6 the results of the case study are discussed, providing an answer to Q4. The most important challenges found are related to privacy, business understanding and data quality/quantity. The results of the case study could provide insights in the experiences of King's Landing. There are lessons learned in King's Landing that can be relevant for other organisations. The lessons learned and generalised conclusions are described in chapter 8.1 and the

interpretation of the results leading to these conclusions are described in chapter 7. These lessons learned are the answer to Q5.

8.1) Lessons learned

The case study at King's Landing has given insights in the opportunities and challenges organisations face when doing data science. Based on these insights, several generalized conclusions could be drawn.

1) **Under the current circumstances, it is unlikely that data science will change the way government organisations create and execute policies.**
2) **Applying data science is especially difficult for government organisations due to uncertainty about privacy laws.**
3) **Organisations should be somewhat familiar with data-driven decisionmaking before creating predictive models.**
4) **Organisations should tackle issues with data quality, quantity and isolated data silos before starting with data science.**

The relation between the 4 conclusions is described in figure 7.

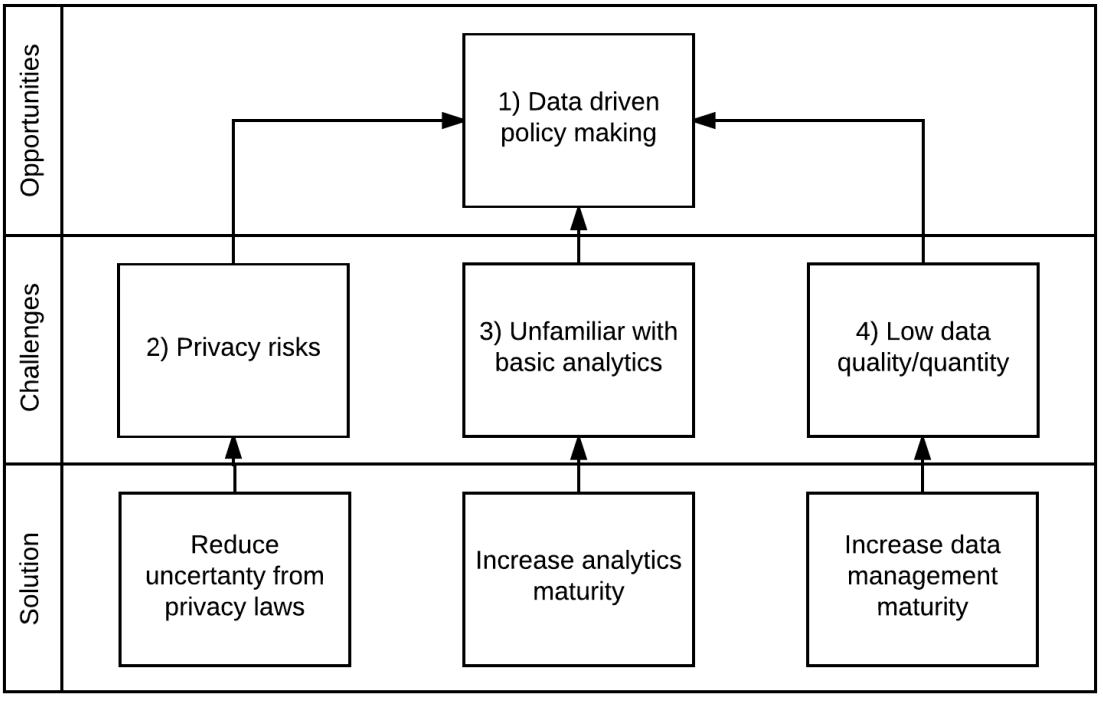

Figure 7: Relations between different conclusions.

Data-driven policymaking is not possible at this time because of the privacy risks, and because the organisation is unfamiliar with basic analytics. In the discussion (chapter 7) the relation between these two challenges, and process and analytics maturity is described.
In the remainder of this chapter each of those conclusions will be elaborated and related to the case study of King's Landing.

Data driven policymaking

1) Under the current circumstances it is unlikely that data science will change the way government organisations create and execute policies.

One of the most important opportunities for the municipality in general is using data science to create and evaluate policies in a new way.

This new way includes the following advantages:

- Less assumptions when creating policies
- Insights in citizens priority
- Faster and more detailed feedback on new policies
- Better implementation of policies

The details of these advantages are explained in section 6.4. Policies are one of the most important mechanisms by which municipalities operate.

Changing the way of creating and executing policies is a complex, long term change process. All the challenges found in the case study would need to be solved. The most important challenges found in King's Landing are the uncertainty about privacy laws, data quality/quantity and the analytics familiarity. Creating policies is a complex process and data science could change it but at this time there are still many challenges. Besides the three challenges mentioned, there might be other challenges which have not yet surfaced.

Privacy laws & government organisations

2) Applying data science is more difficult for government organisations due to uncertainty about privacy laws.

Government organisations are more conservative about data science and privacy than commercial organisations. When there is uncertainty about what is allowed, commercial companies tend to start with data science despite the uncertainty. Government organizations have no pressure to stay ahead of competitors, and fear getting negative publicity. These factors makes successfully finishing data science projects more difficult for government organisations.

Privacy was a big challenge in the case study in King's Landing. New laws regarding privacy are not clear about what is allowed and what is not. Any employee can use the privacy argument to block progress in a project they are involved in. This is sometimes based on their own motives to stop the project. Some of the employees described that the organisation is paralyzed by fear of these uncertainties in privacy laws.

Despite the current culture there are some projects that are being done to get experience with data science. They have taken several measures to ensure that the privacy of the citizens is being guaranteed. At the same time they are attempting to create a privacy framework together with a law firm.

Doing data science in a government organisation is more difficult due to the uncertainty about privacy laws. When more government organisations get experience with applying data science this difficulty might be resolved.

Creating data-driven decisionmaking capabilities

3) Organisations should be somewhat familiar with data-driven decisionmaking before creating predictive models.

Starting data-driven decisionmaking with predictive models is like learning to ride a bike before you can walk. In the analytics maturity models described in this thesis it is advised to use data to gain insights in the past before trying to look in the future.

The first level often entails use of reports and dashboards to support the decisions. When this maturity level is achieved data sources are already linked and prepared for analysis. The people in the organisation are also more familiar with both the data and how to interact with the analytics department.

The stakeholders were not able to come up with research questions with the data scientist. It is the task of the data scientist to assist the business with this, but it might be more effective to prepare the business a bit better. Knowing the limits and possibilities of data science can help the business in asking the right questions. With self-service Business Intelligence users can use tools to find the data they need to support their decisions. When they have a session with the data scientist to come up with research questions for predictive models they will have some experience in operationalizing high level questions.

When stakeholders have little analytical experience it is more difficult to find good research questions. Stakeholders can get analytical experience when their organisation offers them tools and data associated with lower maturity levels. Therefore the lower maturity levels are vital to the success of data science.

Data infrastructure issues

4) Organisations should tackle issues with data quality, quantity and isolated data silos before starting with data science.

The issues related to data quality and quantity can be explained by a lack of data management maturity. Organisations have been using data to support their business for decades, making the field of data management more developed than data science. Data scientists are jack of all trades when it comes to data related topics, but most of them use ad hoc methods to get results quickly. But long term solutions for data management is often not in the skillset of data scientists. Specialised IT people are more efficient in performing these data management tasks.

The data in the JGZ project in King’s Landing is still mostly in silo’s which caused difficulties related to data quality and quantity. If King’s Landing had more mature data infrastructure in place there would be much better data understanding in the organisation. Centralizing data infrastructure is an important component of data management maturity.

Having mature data management is much more efficient than doing data management tasks during a data science project. Data management should give insight in data quality and quantity. It also aims to remove isolated data silos in the organisation, by providing analysable data to the organisation. This data can be directly used for decision making or further processed in a data science project.

8.2) Recommendations

Having a suitable analytics maturity helps organisations when starting new data science projects. But organisations cannot change their maturity overnight, so how can organisations increase their maturity while also starting data science projects?

Data science diffusion in King's Landing is still limited to the innovators, and starting to propagate to the early adopters. One of the goals of King's landing is to increase the data-driven decisionmaking in the entire organisation. The best way to increase the adaptation is to take measures for the different consumer groups in the organisation.

Reduce uncertainty on privacy laws

One of the requirements of acceptance for other consumer groups is that the uncertainty around privacy is reduced. The privacy framework is a good step King's Landing has taken to achieve this.

Create showcase projects to educate the organisation

A showcase project can be done with the input of the innovators of the organisation. The goal of this project should include the deployment phase so actual business value can be generated. These projects can then be used to show other consumer groups the benefits of data science, and how citizen privacy is safeguarded.

Provide training for business users involved in data science projects

Communication between data science and business users is not always trivial. Results from the case study show that forming useful research questions is sometimes hard. It is the data scientist's responsibility to better understand the domain and to involve and educate business users on data science. But the meetings are much more effective when they can be focussed on the problem instead of the basics of data science.

Giving a basic training about data science to business users should make the business understanding phase in future projects more effective.

Stimulate data-driven decisionmaking among all user groups

Starting data science before the lower analytics maturity levels are achieved means that the lower levels need to be achieved in parallel. Lower levels include the use of historical data in reports and dashboards to support decision making. There are some very user friendly visualisation tools that facilitate the creation of interactive dashboards. These interactive tools can replace traditional static reports and they give more insight.

The early and late majority can be tempted to use these self-service BI tools in their decision making. This can be seen as low hanging fruits in the journey to becoming more data driven. Letting them get used to interactive information sources will provide a solid foundation to be involved in data science at a later stage.

These recommendations are done based on the case study results. They are applicable to organisations with challenges similar to what King's Landing experienced.

References

Azevedo, A. I. R. L., & Santos, M. F. (2008). Kdd, semma crisp-dm: a parallel overview. IADS-DM.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance?.

Burton, B. (2009). Toolkit: Maturity Checklist for Business Intelligence and Performance Management. Gartner Research.

Cao, L. (2016). Data science: nature and pitfalls. IEEE Intelligent Systems, 31(5), 66-75.

Carter, D., & Sholler, D. (2015). Data science on the ground: Hype, criticism, and everyday work. Journal of the Association for Information Science and Technology.

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.

Drew, C. (2016). Data science ethics in government. Phil. Trans. R. Soc. A, 374(2083), 20160119.

Eckerson, W. (2007). TDWI Benchmark Guide: Interpreting Benchmark Scores Using TDWI's Maturity Model. TDWI Research, 3-14.

Harris, J. G., & Mehrotra, V. (2014). Getting value from your data scientists. MIT Sloan Management Review, 56(1), 15.

Khan, Z., Anjum, A., & Kiani, S. L. (2013, December). Cloud based big data analytics for smart future cities. In Proceedings of the 2013 IEEE/ACM 6th international conference on utility and cloud computing (pp. 381-386). IEEE Computer Society.

Maoz, M. (2013). How IT Should Deepen Big Data Analysis to Support Customer-Centricity. Gartner G00248980.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. The management revolution. Harvard Bus Rev, 90(10), 61-67.

Mecca, M., Young, R., Halcomb, J. (2013). Data Management Maturity (DMM) Model. CMMI Institute

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decisionmaking. *Big Data*, *1*(1), 51-59.

Rose, D. (2016). Data science : Create teams that ask the right questions and deliver real value. Berkeley, CA: Apress. doi:10.1007/978-1-4842-2253-9

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, *1*(2), 85-99.

Viaene, S. (2013). Data scientists aren't domain experts. IT Professional, 15(6), 12-17.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly, xiii-xxiii.

Appendix A: Interview questions

Phase 0:

- Wat is je positie in de organisatie?
- Wat is je rol binnen het project?
- Welke kennis breng je naar het project?

Phase 1:Ideeen over kansen/uitdagingen

- Welke kansen zie je voor data science binnen je organisatie?
- Welke uitdagingen zie je voor data science binnen je organisatie?

Phase 2: Uitdagingen in current project

Welke uitdagingen kwamen jullie tegen in elke stap van CRISP-DM?
Wat deden jullie, wat ging goed, wat ging minder goed

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

Wat is je hoop en zorgen voor dit project?

Phase 3: comparison with literature

Bespreek elke kans/uitdaging en vraag of ze ze herkennen
Data driven decisionmaking
Challenges in starting data science projects

- Pitfalls about data science concepts (Cao, 2016)
- Overfocus on technology(Rose, 2016)
- Talent management (McAfee, 2012)
- Technology gap with existing IT (McAfee, 2012)

Data science team dynamics

- Nurture Versatile Employees (Viaene, 2013)
- Produce Working Solutions (Viaene, 2013)

Company mindset

- Working without objectives (Rose, 2016)
- Data-Analytic Thinking (Provost & Fawcett, 2013)

Privacy & ethics

- Ethics framework (Drew, 2016)
- Privileged access to data (Carter & Sholler, 2016)

Phase 0:

- What is your position within the organisation?
- What is your role in the project?
- What knowledge do you contribute to the project?

Phase 1:Ideas about opportunities and challenges

- What opportunities do you see for data science within your organisation?
- What challenges do you see for data science within your organisation?

Phase 2: Challenges in current project

What challenges did you encounter in each of the step of CRISP-DM?
What did you do? What went well? What didn't went well?

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

What are your hopes and worries for the project?

Phase 3: comparison with literature

Discuss each challenge/opportunity and ask if they recognise the challenge.
Data driven decisionmaking
Challenges in starting data science projects

- Pitfalls about data science concepts (Cao, 2016)
- Overfocus on technology(Rose, 2016)
- Talent management (McAfee, 2012)
- Technology gap with existing IT (McAfee, 2012)

Data science team dynamics

- Nurture Versatile Employees (Viaene, 2013)
- Produce Working Solutions (Viaene, 2013)

Company mindset

- Working without objectives (Rose, 2016)
- Data-Analytic Thinking (Provost & Fawcett, 2013)

Privacy & ethics

- Ethics framework (Drew, 2016)
- Privileged access to data (Carter & Sholler, 2016)