Radboud University Nijmegen

Faculty of Science

Data Governance in Advanced Analytics: Opportunities and Challenges

Addressing the data governance needs and challenges for Advanced Analytics models

Thesis Msc Information Sciences

Author:
Ton Kunnen

Supervisor:
Stijn Hoppenbrouwers

Second reader:
Arjen P. De Vries

August 2022

Abstract

With the arrival of the data age, new technologies such as Advanced Analytics provide extensive business opportunities for creating value from previously unused data. However, there are also new challenges in how to properly handle this data – and the people using it. Data governance traditionally assists in handling these challenges, but current techniques may not properly translate to Advanced Analytics and the way in which it uses data. This thesis explores these newfound challenges surrounding data governance on Advanced Analytics, and aims to point out solutions wherever possible. It does this through researching existing literature from both subjects, and an interview with an expert in this field to validate the finding from the literature study. This culminates into a governance framework, intended as a general template for data governance use-cases on Advanced Analytics.

Contents

1 Introduction

The landscape of data processing has vastly changed over recent years; technologies like big data, Artificial Intelligence and the Internet of Things have altered the face of data collection and processing. This has led many to describe this period as the data age, with data as the oil of the 21st century. At the core of this transformation is Advanced Analytics, an umbrella term for technologies that can automatically extract patterns and predictions from large amounts of data. In conjunction with this rising importance of big data and Advanced Analytics, the amount of data collected by large companies and the ease with which data is shared has also changed significantly. As a consequence, many new doors in terms of business possibilities are opened. Advertisers can more directly target clients, banks can more easily detect fraud than ever before, and management can process more data to make better informed decisions. Business research suggests that technologies like Artificial Intelligence could deliver an additional 13 trillion dollars per year in global economic output [1].

However, these changes also present new challenges, requiring businesses to adapt their processes. Among these challenges is how to handle all of this data, ensuring it actually creates usable business value, while simultaneously not violating any laws the organisation may be bound to. This is where the field of data governance comes in. Data governance entails all the decisions and policies made to ensure that the data management is handled properly (this will be further discussed under "Introduction to Data Governance"). Without proper data governance, handling large amounts of data over different lifecycles, company sections and processing cases can become an impossible task. And with governments and unions increasingly clamping down on the regulatory freedoms that organisations which process data have previously enjoyed [2], the need for proper data governance is more urgent than ever[1].

This thesis aims to get a grasp on the unique challenges and opportunities that are presented to organisations who are increasingly using Advanced Analytics to create business value, and how data governance can be used and adapted to smoothen out business processes, avoid pitfalls and help in this value creation process. To answer this question, the research question is as such:

"What alterations need to be made to classical data governance to suit the needs of data governance in the context of Advanced Analytics?"

To get a clearer grasp on these overlapping worlds, this main question is separated into two sub-questions:

"What are the challenges currently present within data governance on Advanced Analytics?"

This question mainly focuses on what literature studies on current data governance show, and how these can be translated to Data governance on Advanced Analytics.

[1] Despite its necessity, data governance is only a piece of the puzzle in organisations moving towards a data-driven style of analytics. A data-driven culture, support from corporate sponsors and a proper understanding of the AA tools are also key to ensuring this transition is as smooth as possible. However, data governance is the bedrock upon which all of these transitional stages are built; without proper data quality, access and performance verification, none of these advanced technologies are capable of creating value.

"What possible solutions exist for the challenges currently present within data governance applied to Advanced Analytics?"

This question builds on the first sub-question by focusing on how to approach the challenges presented there. Although it partially answers some of the mentioned challenges, it also casts a wider net, by including general solutions that may not directly fit a single mentioned challenge. Instead, these aim for general improvements towards data governance models, as to function better overall.

2 Methodology

2.1 Outline

At its core, the subject of this thesis lives on the edge of two fields that do not often overlap in the research world; business and technology. Due to this multidisciplinary nature, an extended theoretical background to both data governance and Advanced Analytics is provided, in order to accommodate readers from different backgrounds. This includes an in-depth look on subjects such as Advanced Analytics algorithms, relevant pieces of law surrounding data protection and regulation, and a deeper understanding of data governance itself. This section also serves to introduce some of the concepts found in the literature study, upon which the rest of the thesis builds.

Following this section, the "Challenges and Opportunities" section provides an overview of possible challenges derived from the literature study that may affect future data governance projects using Advanced Analytics. When found, possible solutions are posed towards these challenges.

The Interview section expands on the findings from the previous section, by discussing how someone with experience in data governance sees these challenges within their organisation. Since the literature study is mostly done from a theoretical background, this section should help ground the information found here into a real business case, and see whether it translates to real life.

The thesis closes with the Conclusion and Discussion sections. The Conclusion sections aims to summarise the findings from the thesis, and present a proper answer to the research questions posed. Lastly, the Discussion sections aims to discuss the findings summarised in the conclusion section. This includes discussing the overall results of this thesis, and interpreting the findings from the interview. It also looks at the limitations that this thesis ran into, and how future research can avoid these limitations and expand upon the findings from this thesis.

2.2 Literature Study

A large focal point of the starting stages was the literature study, which was a challenge in and of itself; the subject of data governance in the context of Advanced Analytics is rather new, and not of the interest of many data scientists and analytics researchers, as it doesn't directly affect system performance. This means that little is written about it (more on this under "Cultural Resistance"). Thus, much of this section is of an exploratory nature. Conceptually, the literature study is divided into three sections:

1. The data governance section: This phase focused on grasping the basics of data governance, commonly used structures and how it is used in non-Advanced Analytics environments. The importance of this phase was mostly to capture the underlying decision process of data governance, as it is important to understand when making adjustments to it for the sake of Advanced Analytics.

2. Advanced Analytics section: This phase focused on understanding the components of Advanced Analytics relevant in the context of the processes described in this thesis. Unfortunately, the entirety of Advanced Analytics itself is too wide, dense

and ill-defined of a subject to fully discuss here - not to mention out of the scope of this thesis. This means that not all elements of Advanced Analytics are included in the literature study, and instead the focus shifts to the components that are relevant for applying data governance.

3. Other: This phase discusses various subjects that are relevant for data governance in organisations. This includes relevant legislature (such as the GDPR [2]), societally relevant things such as discrimination, and general organisational factors, such as the culture of an organisation. This phase also discusses subjects relevant to glueing together the data governance and Advanced Analytics phase, as this is a core component of this thesis.

2.3 Interview

Despite the value of the literature study, most of the concepts in this thesis arose from reasoning about existing theories and concepts, not actual implementations. To test whether the concepts formed during the earlier stages of the thesis are grounded in reality, an interview with Rene Snijders was conducted. Mr. Snijders has ample experience in the field of Business Intelligence and Data, working at Nuon and Vattenfall in Data management and BI before moving to Alliander as Chapterlead Data & Analytics. This is a Dutch network operator company, operating the gas- and electricity network at both low- and medium voltage.

The interview mainly functions as an extra qualitative component, more specifically as an addition to the literature study. It mostly serves to validate the results found in the literature study. It does this through taking the challenges and possible solutions from this thesis, and validating these against Mr Snijders' experiences in the field. As his experience in the field is both extensive and deep, this provides some valuable ecological validity to this thesis.

The interview has a semi-structured form to maintain flexibility in responding to answers, while ensuring the interview remains focussed. This is considered a good fit for research of an exploratory nature, as it allows for new and unexpected concepts to pop up, while enough structure remains to guide the conversation. Prior to starting, permission has been requested to record the interview. Processing-wise, a non-verbatim translation process has been used, which was then turned into a more readable section in the "Interview" section. Apart from that, as little as possible has been altered to the original text, with translation occurring as late as possible. This is done to maintain the original intentions of the interviewee as well as possible. The interview was conducted in Dutch, with all embedded quotes translated as directly as possible.

In terms of qualitative analysis, no coding tools have been used. As only one interview has been conducted (and the value of coding mostly comes from multiple interconnected interviews), this would add little to the overall conclusion. Instead, the focus is on how the answers from the interview line up with the elements discussed in the earlier sections of the thesis, and whether there are any omissions or additions to that content.

2.4 General Information

In terms of mode of argumentation, this thesis leans mostly towards abduction. As the Field of data governance on Advanced Analytics is still quite young, it lacks the premises to fully reason using deductive and inductive argumentation. Instead, we are looking for a conclusion (in our case, solutions and challenges) that best fits an observation (in this case, the literature study). Although this form of inference is considered the least accurate of all three, it best fits the current structure of this thesis.

As an end product, this thesis presents a governance framework that could apply towards organisations that use Advanced Analytics. This is by no means a final or universal product, but it can serve as a guideline for those who are setting up a data governance effort that includes Advanced Analytics in their own organisation.

2.5 Validity

One aspect of exploratory, qualitative research that is rather hard to define is that of validity. This is due to the fact that despite strong considerations on the content written here, the lack of quantifiable results and a testing environment make it hard to measure the outcomes of the concept described. Here, we are mostly referring to internal validity, which concerns itself with how much the observed effect can be produced [3]. Due to the aforementioned theoretical nature of this thesis, statistically proving validity is unfortunately not an option. Thus, our method of validation mostly leans on construct validity. This method is mostly focused on how well the research measures the construct, which is done by observing related indicators [3]. This is synonymous to how a disease (the construct) has symptoms (indicators) indicating its existence and impact. In our context, this refers to how factors like time loss on data access requests, frequency and impact of data breaches and storage costs indicate the existence of poor data governance.

2.6 Relevance

At its core, not all issues surrounding Advanced Analytics models can be solved through data governance. For example, the Black-box problem (see "The Black-Box Problem") is a problem inherent to the technical workings of many Advanced Analytics models, and will not be solved through a business angle. However, proper data governance can help in mitigating and reducing the impact and likelihood of issues occurring. When combining Advanced Analytics' downsides (e.g. the opaqueness of these systems) with poor ideas and implementations concerning data accountability and governance, minor situations that could be handled through proper data governance now risk expanding into larger problems that can cause legal- and trust-issues for an organisation. And although many of the current technical issues may be solved through technological advancements, the necessary responsibilities towards data usage from a business perspective remain relevant, and require well-thought out solutions for things to run smoothly.

Another advantage from approaching this issue from a business perspective rather than a technological perspective is that it tends to apply more universally across technologies. For example, a technological solution that improves the interpretability for one type of Advanced Analytics model is unlikely to directly transfer to a different model. In contrast, data governance from a business perspective is based around the structuring

of organisations, which exists regardless of the chosen Advanced Analytics model. This means that data governance solutions are likely to be more resistant to changes over time and different technologies than technical solutions.

3 Fundamental Concepts and Background

3.1 Introduction to Data Governance

Data governance is a term that is not universally defined, but the definition that is most likely to fit some universal need comes from the Data Management Body of Knowledge, commonly referred to as the DMBOK [4]. This describes data governance as:

"The exercise of authority, control, and shared decision making (planning, monitoring and enforcement) over the management of data assets."

An important distinction here is that it is discussing governance over the management, not the management itself. Although the fields are closely aligned and are likely to overlap, their end goal is not the same. Data management focuses on the actual implementations used to create value from the data (i.e. how to handle the data itself), while data governance focuses on ensuring this management is actually executed correctly. This is generally expressed in the V, a framework which separates the functions of data governance and data management (see Figure 1). This separation is further discussed under "Rethinking the Separation between Management and Governance".

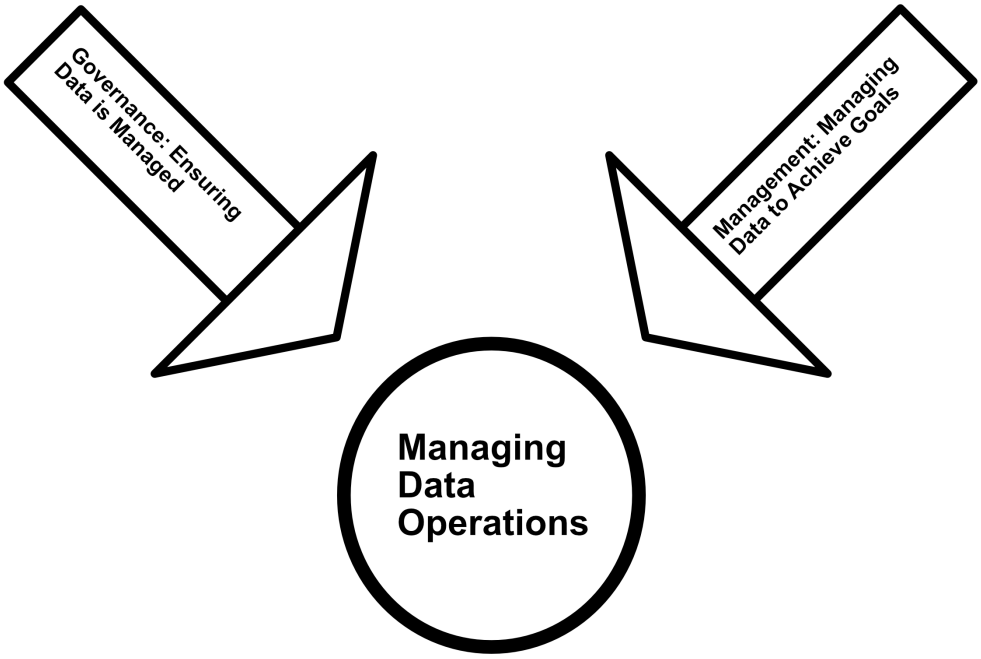

Figure 1: A functional view of "the V", adapted from [5, p. 122]

A well-fitting metaphor is that of policing the quality of data within an organisation. Although terms like "policing" tend to bring up negative connotations, data governance does not aim to be a restrictive, bureaucratic process [6] (although it often ends up being one). Instead, it aims to help organisations work better together with their data and each other, by ensuring that the processes surrounding this data and their users are well-managed. For data to be transformed into valuable information, it has to be accurate and trustworthy. How organisations handle this differs per organisation and

their respective data (and the information they aim to extract from this data), but every organisation has to ensure the quality of their data before its data value can exist.

This auditing process is not something all too foreign to most organisations. Although audits are historically associated with financial audits of the financial accounts, many facets of an organisation can be audited. Examples are ensuring that HR complies with labour laws, whether vendors still meet the standards that organisations require of them, or whether IT security meets standards set by international commissions (such as an ISO 27001 audit [7]).

Apart from complying to legal standards, audits also provide value for the organisation itself; either it receives a validation that ensures it meets some recognized standard (which provides value to customers), or it gets a clear view of any failures in the organisation (which provides an opportunity for growth and improvement).

One important thing to note is that data governance is not so much about the governing of the actual data itself (even though it generally ends up as a facet of it in some way). Instead, it mostly concerns itself with the governing of the people interacting with the data. As John Ladley [5] puts it,

"It is about deciding what people can and can't do with data, as well as ensuring that there are guard rails in place to make that happen."

Although data governance often feels like a new concept that emerged in the wake of big data's rise, it is actually something that has been done since the inception of most data systems [5, 6][2]. After all, the need to manage data is almost inherently tied to the reasons for having data in the first place. One of the earliest modernised forms of data processing was during the US Census in the late 1800's, when it became clear that manual processing would take too long, and that databases would only keep growing in size [8]. Bureau engineer Herman Hollerith – who ended up founding IBM [10] - employed a punch card system to assist in the management of the census' data processing.

From an outsiders' perspective, many of the discussions and challenges surrounding data governance may seem overly complex and unnecessary, since most people have never had to explicitly manage and audit their data. After all, the average persons' data is limited in scope and handled personally. However, there are various reasons to do data governance:

- Regulations such as the GDPR may require a per application review of data requests, which becomes impossible to maintain in large organisations unless it is properly managed and audited. Sometimes the transparency of the data flow is also important, requiring clear documentation of who was able to access what data and for what reasons.
- It is important to divide responsibilities surrounding data governance so that people can be held accountable for their decision. However, data analytics often works with long-term projects that have many employees interacting with the data, while also creating new data. This is where a lot of the value creation occurs, but it also makes it harder to maintain that responsibility and accountability. Data governance provides the structure to maintain this responsibility and accountability.

[2] A case can be made that data governance goes back even further, such as the Ishango bone from 18000 BCE [8] or the earliest Egypt census [9]. However, this carries little relevance for this thesis and will therefore not be referenced further.

- Many organisations handle some sort of sensitive data, even if it is not part of their value creation process. For example, employee information created by HR can damage employees (and the company) when it falls into the wrong hands. Data governance can ensure the protocols surrounding this data are executed properly.
- For some branches of organisations, the validation from data governance itself already provides value. For example, finance departments are likely to heavily value the certainty and consistency of proper Master Data Management (see "Master Data Management on Stochastic Systems section"), as this legitimises their efforts.

There is a pretty clear separation here in terms of internal and external antecedents. Most internal antecedents point in the direction of business process harmonisation, where the focus is on aligning business processes through usable access to data that is properly maintained. Large amounts of time are spent by data engineers on finding the proper data and ensuring it is correct, instead of processing data to create value for the organisation. Recent surveys suggest that data scientists spent 45% of their time on data preparation tasks such as loading and cleaning data [11] - a number that rises to 80% in some surveys [12]. This is a very inefficient use of resources, as these preparations detracts from any actual processing being done.

Externally, the focus is mainly on meeting legal and regulatory demands placed upon the organisation, to avoid any fines and bad publicity.

3.1.1 Introduction to Data Governance roles

Setting up a data governance effort involves a myriad of factors that have to be considered, from governance structures to policies surrounding the use of data assets. First, let's present a basic role structure for a general data governance model. Although there is no completely universal structure of data governance, this role division is generally present in literature [13], and is likely to translate well towards most organisations. From an organisational standpoint, the modern data governance structure is generally separated in three main roles (definitions adapted from [14, p. 60]):

- Data owner: This is the person ultimately responsible for a data set. This person ensures that data is fit for the purpose of the data user(s). Ownership is generally given to the creator of the data asset, as this allows for correctness verification.
- Data user: This is the person that wants to use the data. The data user typically negotiates with the data owner about data access. These negotiations contain subjects like what data needs to be used, what the data definitions are, or what the possible data quality requirements might be.
- Data steward: This is the person with hands-on responsibility for managing the data. These tend to have a mixed business/IT background.

The core interaction here is the negotiation process between the data owner and data user (see Figure 2). The data user creates a request to access the data, which is received by the data owner. Due to the data owner's knowledge about the dataset and its correctness, this person is also the gatekeeper for access to the data user, and aims to ensure that access towards this dataset is handled properly. The steward assists in the interaction between the other two roles.

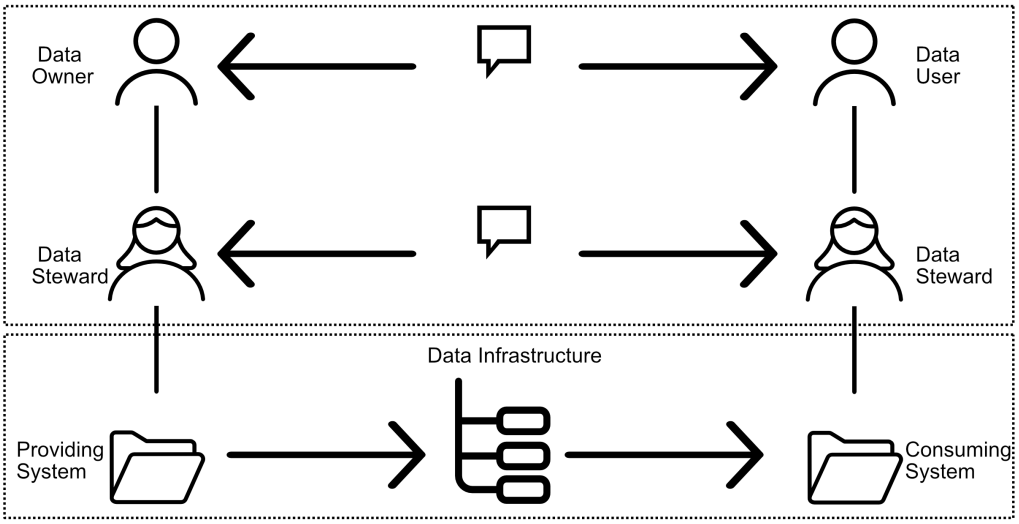

Figure 2: A general data governance structure, adapted from [14]. The upper section represents the interaction between the user, owner and steward, while the lower section represents the data storage flow

Normally, activities and responsibilities within projects and organisations are captured in RACI-matrices. The aim of these matrices is to present clarity around the roles and responsibilities for each project participant and stakeholder [15]. These matrices place personnel roles on the columns and activities on the rows, with RACI representing their role and responsibilities to these tasks; one is Responsible for the task, Accountable for the completion, Consulted, or Informed about the progress. Especially in the context of data governance, RACI-matrices are useful for clarifying the role separation.

3.1.2 Current State of Data Governance

Although many companies see the potential in the power of data, most are woefully unprepared for the challenges that arise when handling this data.

After conversations with an external organisation, it became clear that there is still much room for improvement within this field. Even though this organisation has already invested large amounts of time and effort into improving their data governance, they still run into its limits on a daily basis. For example, they are currently working in a model where the role of data usership – the role which determines what data is required for a specific use-case – is still undefined. Nevertheless, this role still exists and is relevant within the organisation, even though it is not formalised. This in conjunction with the need for data owners to often approve data access on a case-per-case basis, leads to a highly inefficient system for data governance; even though their current system covers legal requirements and possible internal transparency goals, it eats up a lot of time for data owners. These are not exclusive roles to these employees; data owners are often full-time employees who have been told to fit any governance tasks into their working time. Apart from the added workload, this can lead to enormous delay in data collection (and thus processing), and leave projects stuck in approval-limbo; a 6 month gap between data request and approval is not unheard of.

Many of the problems seen now will only worsen when we see the next big wave of data processing models grow in usage, often referred to as Advanced Analytics.

3.2 Introduction to Advanced Analytics

Up to recent times, most data (and thus data governance) used in organisations was processed in the context of Business Intelligence. This method aims to take the data found in organisations, then analyse and present it (e.g. through dashboards) in such a way that it can help managers make more accurate and informed decisions about their organisation. However, recent years have seen a shift in this status quo, with the rise of Advanced Analytics.

Advanced Analytics (sometimes referred to as AA) is a rather broad term, with little in terms of an official definition. One characteristic that is considered relevant to many AA techniques is that it extracts high level data - such as patterns and optimizations - from low level data such as texts, images, sensor information or social media data [16]. In general, it's used as a blanket term for high-level data analysis tools and techniques. It encompasses (but is not limited to) technologies such as big data, Artificial Intelligence and data mining.

Despite appearing similar, most businesses tend to make a distinction between use-cases for Business Intelligence and Advanced Analytics. Where Business Intelligence is generally used for getting insights from data to explain previous occurrences in an organisation (essentially properly summarising inputs to create advice for future references), Advanced Analytics is geared towards predictions and automated recommendations [17].

Although there are many types of Advanced Analytics models and approaches, most of them boil down to one of two categories[3]:

- Classification: Classification aims to categorise objects in a dataset based on some characteristic of the objects. It does this by learning the characteristics of an object group through training on historical data, then categorising new data based on this training. An example of this is a spam filter, which classifies incoming e-mails based on whether they contain legitimate content or not. The use-cases for this type of Advanced Analytics are varied, and range from product quality assessments by classifying by quality bracket, to identifying customer behaviour based on purchase characteristics. A close relative to this is clustering, with the main difference that the categories to which elements are categorised are not predefined in clustering.
- Regression: Regression takes existing patterns from previous occurrences, and attempts to predict future behaviours based on these occurrences. A classic example is organisational growth, which is often a prediction of the future based on classical data. One use-case is predicting new markets to invest in based on their projected growth.

These different systems are often combined into a full data chain. For example; using classification on a customer database to create categories based on purchasing behaviour, then using regression to predict future purchases for these categories.

[3]Unsupervised learning contains additional categories such as Principal Component Analysis and dimensionality reduction [18], but these fall outside the scope of this thesis.

In general, two main advantages are presented in favour of using Advanced Analytics over classical Business Intelligence. First is the possibility to actually create meaningful outputs from large amounts of unstructured data. As it stands, most organisations do not create significant value from the majority of their stored data (this will be further discussed under "ROT/Dark Data"). There are a multitude of reasons for this, but the most important is that classical analytics tools simply do not have the capacity to handle these large flows of unstructured data in any valuable sense. With large organisations creating data flows of up to 4 Petabytes per day [19], aiming for classical analytics is simply not viable. It also means that organisations spend large amounts of resources on saving assets that provide zero value, something no organisation would want to be found guilty of if it pertained to any other asset. Worse, organisations might lose grip on the data that they do have, and run into accidental illicit uses (for example, sensitive unstructured data being stored or used beyond what is legally permitted, creating legal disputes and facing image damages). With Advanced Analytics, one is able to use these large swaths of unstructured data to support the main business purpose - actually extracting value from it.

Second is the real-time processing capabilities that many Advanced Analytics technologies provide. The new age of data is not only defined by the quantity of data that is gathered, but also the speed at which it enters the organisation. With web crawlers and sensors providing a constant stream of high-speed readings, the time-frame of classical BI dashboards and manual upkeep is simply unsustainable. With Advanced Analytics, models are often trained beforehand, making it possible to automatically process data in a live environment while still handling these large streams of data. This opens up the door to real-time processing of raw data, such as classifying objects in a live video stream.

3.2.1 Data Governance over Advanced Analytics

With these new methods of acquiring and processing data, it only makes sense that unique challenges rear their heads that are no issue in classical data governance. For proper data governance in Advanced Analytics, one not only needs to receive the data quickly and efficiently enough to offer any value from the model outputs, one must also have clear and transparent contact with the creator of the training (and testing) data sets, to ensure any results can easily be backtracked through the data. This contact exists such that users and owners can be held accountable, and improvements to data management can be made. These are standards that when not met by organisations, can provide large risks for future projects using Advanced Analytics.

Companies themselves often understand these risks; recent surveys report that 57% or organisations say that they are unable to keep up with the growing rate of data volume [20], with 47% believing their organisation will fall behind when faced with this rapid data growth. A majority of organisations also mention that they do not have the requisite understanding of "Data age technologies"(which the survey defines as 5G, edge computing, blockchain, Alternate- and Virtual Reality, Artificial Intelligence and Machine Learning). Artifical Intelligence and Machine Learning in particular (which cover a large spread of Advanced Analytics techniques) show lacking numbers; 42% of IT and business managers did not have an expert understanding of these technologies, despite 51% reporting that their organisation will be using these technologies in the future. It is clear that despite the enormous potential of these systems, most organisations are not ready for the transition to this new form of data processing.

An added issue is that since most companies are still relatively new to the use of Advanced Analytics technologies like Artificial Intelligence, they have not yet had the time to get a tacit grasp on how many of these technologies should be used, or hone their instincts around them [1]. Issues that show themselves surrounding societal or organizational risks often require time to get a working knowledge of and understand their potential dangers, leading to companies using these systems in places they may not be suited for (for example, automated decision making in functions with an increased risk of discrimination). It can also lead to underestimating the specialisation required, either believing that their standard risk-mitigation techniques will also cover any Advanced Analytics-related issues, or that their general IT team can properly handle all of its unique quirks and characteristics.

3.2.2 The 5 V's of Big Data

As Advanced Analytic often relates to big data, it is also important to discuss its factors in the context of data governance. Big data has a few dimensions that are relevant in how it's governed [21, p. 431]:

- Variety; this refers to its structure, which can be structured, semi-structured or unstructured.
- Velocity; this refers to the high processing rate, which is important for the aforementioned real-time processing of data by Advanced Analytics.
- Volume: this refers to the high growth rates of big data, which is inherent to big data; many different flows – all with constant data inputs – create a high growth rate which may require adjustments to data governance to safely process.

More recent versions expanded on this with 2 more factors[4] [22]:

- Veracity: This refers to the accuracy of the data, and more importantly, how much trust one can have in any outputs resulting from the data.
- Value: This refers to the extent to which value can be derived from the data. For organisations using big data, this factor is likely to be a culmination of all the other factors, as all present some information about how valuable the data can be for a project.

With these 5 factors, the challenge that big data presents over classical data also becomes fivefold:

- Volume: Whereas data could previously be processed in a reasonable timeframe, companies now often find themselves unable to process the data in a timely manner due to a lack of resources. Thus, processing gets deferred, leading to a build-up of dark data [23].

[4] Although some sources include additional factors such as Variability and Visualisation, these are not consistently used in literature and are as such not mentioned from here on out.

- Velocity: Due to the high speed of real-time processing, the timeframe of processing is not only shortened extremely, the data itself is also outdated at a far higher pace than previously.
- Variety: With the possibility of unstructured data, there is an increased risk of losing control over the contents of your data, compared to fully structured data.
- Veracity: With the large amounts of data and the added unstructured state of this data, there is an increased risk of having inaccurate, incomplete or other untrustworthy data, possibly making model outputs unusable.
- Value: As this factor is a culmination of the other factors, and as such, is also a culmination of all the other challenges. It is also the most important one in the end, as poor value extraction from data defeats the purpose of having this data in the first place.

3.3 Introduction to Challenges and Opportunities

Naturally, all of the data governance ideas and implementations discussed in the upcoming sections are highly dependent on what kind of organisations they are applied towards. There are many factors that determine how the concepts discussed below should be implemented, such as:

- Organisation size: Larger organisations may require more effort to implement solutions over the entire organisation, but are also more likely to have the resources required for these implementation efforts.
- Organisation geography: an organisation with an international reach may have to deal with varying regulations and customs surrounding data governance, placing limitations on what can be implemented organisation-wide. For example, the USA lacks any data privacy legislation on the federal level[5], and instead lets states create their own legislation (such as the CPRA [26,27], or the Colorado Privacy Act [28]). This is in contrast to the EU, which has a universal data processing regulation with the GDPR [2]. This means that an organisation may have to create separate protocols and guidelines (or even data governance teams) for every state it operates in[6].
- Organisational data type: Organisations that deal with highly regulated data (e.g. sensitive data as defined in Article 9 of the GDPR [2]) or work in regulated industries may be more capable in asset control, since this is something they have previously dealt with. However, they may also be more restricted in how they can process the data.
- Current IT structure: If the organisation has previously set up a data governance system, it may be required for any AA-specific alterations to fit into this structure for consistency across the company. This may rule out some solutions (for example, a fully decentralised structure for Advanced Analytics may be hard to combine with a federated data governance model). This point also holds for general organisation architectures.

Lastly, one important factor that will heavily affect how solutions are implemented in the organisation is their Information Management Maturity (often referred to as IMM). This factor refers to how capable an organisation is in executing information asset management [5, 29]. Generally, this is expressed through an Information Management Maturity Model. This model scales from the Initial stage – where business rules are non-existent and information maturity is chaotic – to Optimised, where information asset management is woven into the fabric of the organisation. If an organisation has a low IMM, aiming for the most advanced and complex forms of data governance may be something that the organisation is not ready for yet. This factor is closely tied to culture, as the deep integration of data governance requires a cultural understanding of its benefits, combined with a capacity to create changes in the organisation.

[5] Unless one counts federal statutes such as the HIPAA [24] or the Sarbanes-Oxley Act [25].

[6] This is also something where CoE's for separate legislations may become useful, as these allow for a more focused approach toward each relevant legislation(See "Centres of Excellence").

4 Challenges

4.1 The Black-Box Problem 101: A Refresher Course to Neural Networks

One of the industry-specific components Advanced Analytics models often deal with is the concept of black-box systems; systems that are very hard to interpret based on their inner workings, even though they take clear inputs and provide clear outputs. To fully understand the issue at hand, a small crash course on an Advanced Analytics model is useful. For this example, a neural network with supervised learning is used, but as the core issue is the same, the issues presented here persist through all forms of Advanced Analytics to some degree.

Note that this explanation is highly simplified, and ignores many of the elements generally found in neural networks such as convoluted neural networks, stochasticity and unsupervised learning. It mostly aims to explain the main principles of a neural network, and the basics of the black-box problem.

For our example, let's take a neural network that learns to recognize numbers based on images from said numbers (see Figure 3). Such a network can be useful for automatically translating handwritten numbers (such as old banking information) into digital information. As this task revolves around classifying objects (the images) to a certain category (different numbers), it is considered a classification problem. Lastly, we are using supervised learning, which means that during training and testing, inputs and outputs are known in advance; the images are all labelled with their respective number, so that the network can see whether it gave an accurate prediction. This then allows for actual usage of the network when fully trained, without requiring labels.

The basic structure of a neural network only consists of three components; an input layer, a hidden layer, and an output layer (the functions of these will be discussed as we go). Each layer then consists of a multitude of neurons, which execute the core functionality of the network. Lastly, these layers are then connected through interconnections. These are weighted to allow for the values of the neurons to change over time.

For the inputs of this neural network, we take said pictures of images, and convert them into something a network can process. For this case, we are converting the image into grayscale values of every pixel (for example, 0 for a fully black pixel, and 1 for a white one). In this case, every input neuron of the network receives the grayscale value of one pixel of the image. The values of these neurons are known as the activation of a neuron. The activation values of these neurons determine whether a neuron needs to be activated - known as "firing" - or not (or anything in between, depending on some of the mathematics behind the activation).

This idea of neurons turning on is analogous to how actual neurons in the body work; when inputs (either internally or from the environment) occur that are relevant to the function of a neuron, an electrical charge is transmitted, and the neuron "activates"[7]. To continue the biological metaphor, the interconnections between neurons would take the role of synapses, which transfer the electrical signals from one neuron to another. As we have no knowledge about how our network is doing yet, the weights of all the

[7]In the human body, this is known as an action-potential, and is expressed in some form of electrical discharge. As deep and interesting as this analogy is, most of it lies outside of the scope of this thesis.

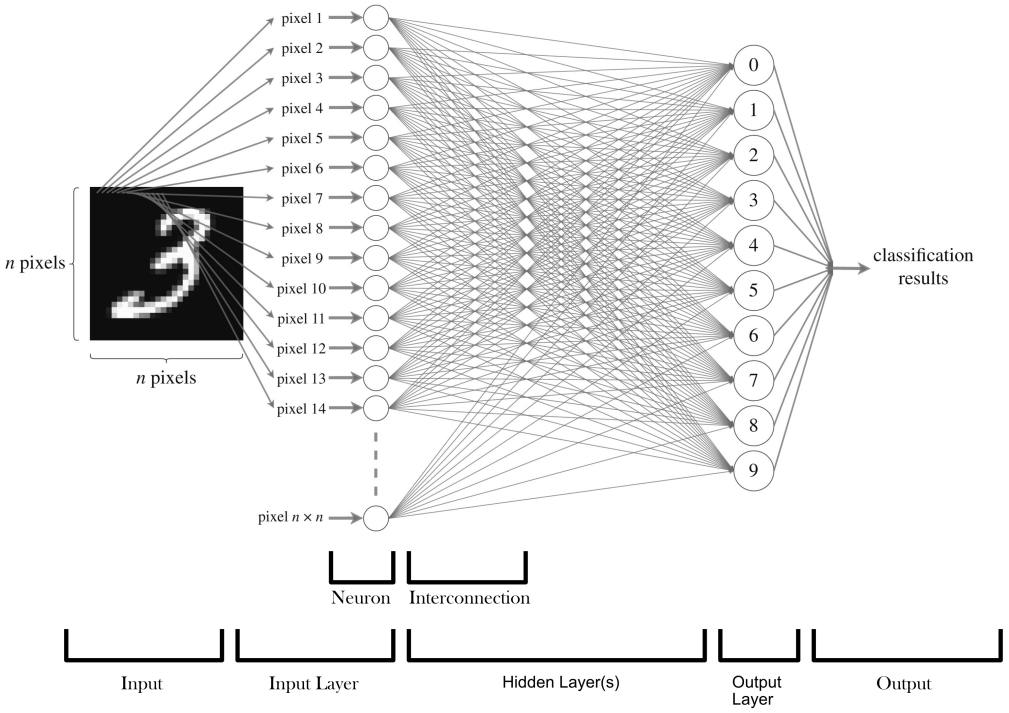

Figure 3: A basic overview from our number-recognition neural network

interconnections are assigned at random.

After applying some smart mathematics[8] on these weights, the interconnections feed their values to the next layer, known as the hidden layer. This is where most of the magic of neural networks happens, and also where the "black box" resides. The process here is essentially the same as what is described above; neurons have a certain level of activation, the interconnections get weighted based on some mathematics, and these weights of the interconnections determine the activation of the next layer. This can either be the output layer, or another hidden layer, depending on your network. This travelling of data from inputs to outputs in the network is what defines this network as a feed-forward network.

After these processing steps in the hidden layer(s), we reach the final conceptual layer of the network; the output layer. This is where all the weights and activations from the hidden layers get transformed into a usable output. In our case, this is what number the network predicts we gave it as input. When trying this for the first time, the network is likely to give a completely inaccurate answer. This makes sense, as very little information has been provided to the network about what it is supposed to do, or how well it is doing it (this is also why the weights were assigned at random at first).

This is where the "training" of the network occurs. As mentioned, the images we put into the network were labelled with their respective number. Comparing this label to the network's output allows the network to be told whether its answer was correct, with an accompanying error rate to see how incorrect the answer was. After receiving this

[8]The weights are multiplied by the input signal, a bias is added, then everything is passed through an activation function.

answer, the neural network alters the weights of its connections (and thus the neuron activations) by going backwards through the network, starting at the output layer and moving towards the input layer. This process is known as back-propagation. As the node activations have been altered, a different answer is provided - hopefully, with a lower error rate. This iteration process continues until the network provides an outcome with an error rate low enough for its use-case. Once this is achieved, the network is ready for testing, after which it can be used in real use-cases.

As complex as this may look, it is essentially an optimization problem; "what are the optimal weights I can give all these interconnections, so that the resulting output gives a minimal error rate?" This parameter optimization is a core component of any AI model, with many different optimization algorithms for different use-cases [30, 31].

When looking at the steps the network undertakes to find its outputs, the core issue of explainability in Advanced Analytics arises. At no point in the processing chain does the network use any reasoning related to the actual inherent meaning of the inputs. Moreover, the network makes no attempts whatsoever at reasoning in a similar fashion to what a human would do. For example, it does not define the characteristics of the number 8 by two stacked, interconnected circles, as any human would. Instead, everything is based on the weights and activations of a network, which themselves are found through trial and error (albeit guided trial and error).

For use-cases of these models where one only cares about accuracy and not how it is achieved, this lack of interpretability is not a direct problem. The issues arise when organisations are aiming to do analytics on the internals of their network[9], or when they are required to explain decisions the network was involved in. For example, if a person asks an organisation after an interview why they weren't hired, the answer "because an algorithm gave these specific weights to nodes" is generally not satisfying (or legally compliant, when faced with accusations of discimination [33]).

The challenge increases when the network is built in such a way that it does not create outputs interpretable by experts. For example, Support Vector Machines (commonly referred to as SVM's) are supervised learning models that are great for classification, especially for very complex analytics problems [34][10]. However, SVM's are also very much black-box systems, and are notorious for their issues with interpretability. They transform their model inputs into vector representation, which generally cannot be interpreted by experts [32]. This transformation of output data creates the difficulties in understanding the model's output, and reduces interpretability. And although there are attempts to increase the interpretability of SVM's [35–37], none so far have been successful in removing interpretability as a factor to be considered when selecting an SVM model.

4.2 Master Data Management on Stochastic Systems

In organisations, there is a possibility that there are multiple, comparable systems creating similar outputs on similar data. This can be useful, as every system can serve as a specialisation of a general concept for some specific purpose. However, when working

[9] It can also create problems when tuning results for accuracy, but this is irrelevant for data governance and thus outside of the scope of this thesis [32].

[10] This is mostly defined by their ability to classify data which is not linearly separable, something a straight line or classical hyperplane is incapable of [34].

with large amounts of data, this can lead to multiple sources of data within an organisation, and conflict about what the "right" data is. This lack of clarity and consistency can hamper performance and accuracy. To handle this, data governance has a concept known as Master Data Management (generally referred to as MDM). MDM aims to resolve these conflicts by creating a "single version of the truth". The DMBOK describes MDM's purpose as [4]:

"Master Data Management entails control over Master Data Values and identifiers that enable consistent use, across systems of the most accurate and timely data about essential business concepts."

This is where Advanced Analytics might provide an added challenge in the form of stochasticity. Contrasting with deterministic systems, stochastic systems use some form of randomness in their processes to create their outputs. For example, the aforementioned number-classifier Neural Network (see "the Black-box Problem") often does its backpropagation step with a technique called stochastic gradient descent [38], which incorporates a random factor to more easily reach an optimal parameter combination. Although this randomness can be very useful - it helps the parameter optimization program by avoiding local minima [38] - it also clashes conceptually with the idea of "the single truth" of MDM. When an organisation uses the same models on the same data yet receives different outputs, it can become hard to match data across the organisation. It can also make it harder to justify the usage of an algorithm; it's a tough sell to say that an application processing algorithm rejected a person based on a random factor (especially in tandem with the concepts discussed in "Discrimination").

4.3 Discrimination

Of all issues surrounding Advanced Analytics techniques like Artificial Intelligence and big data, one that has garnered the most attention in the media in recent years is that of discrimination. At surface level, the idea of discrimination by Advanced Analytics seems odd; how can a system that does little more than transform inputs in outputs be discriminatory? The thing here is that the system itself is not actually discriminatory; rather, it takes data which contains discriminatory biases – whether intentional or not – and transforms them into outputs which are considered discriminatory [39]. For example, when Amazon decided to scrap their automated recruitment system after it showed biases against recruiting women [40], the bias came from the fact that the data the system was fed – the company's applicants over a 10 year period – simply contained the patterns from a male-dominated industry - patterns the AI correctly captured. As the current state of their company was treated as a benchmark to aim for, any characteristic that did not meet their current criteria – among which was being male – was penalised by the system. Even after training the system to avoid these direct anti-female biases, it still found other ways to find and abuse these patterns, such as punishing achievements on resumes primarily done by women.

We see these biases all over AI applications [41]. It also shows why Advanced Analytics is not the holy grail of "computer objectivity" it is sometimes seen as; when these system are fed data that contains that the trials and tribulations of the real world, they will be reflected in the outcomes of the models, as their job is simply to process this data in a very specific way.

The issue of discrimination is also tied to legality, although this depends on locations

and their respective laws. The racial equality directive [42] makes a distinction between direct and indirect discrimination, which is mostly based on intent. Direct discrimination refers to discrimination as most know it; an active effort to unjustly treat someone based on some characteristic, such as rage, age, sex or disability. Indirect discrimination is a bit more complicated; it refers to any apparently neutral practice that harms people with a certain ethnicity [42]. Whether this occurs by accident or on purpose is irrelevant here. For countries bound to the GDPR, both direct and indirect discrimination are banned, unless there is a legitimate aim and the practice is proportionate [42]. Law itself still struggles with these advances, as the nature of these algorithms makes it that AI-driven discrimination may remain hidden (discrimination in general can struggle with this, as it requires a direct comparison to other people who use similar services).

Although there are attempts at removing these discriminatory bias in Advanced Analytics, most of these are focused on altering inputs and outputs, such as data preprocessing to remove implied discrimination, or changing final predictions. One recent direction is that of grey-box ensemble models [43], which uses a white-box classifier model (generally one with good explainability, such as a decision tree) that is trained on the outputs of the original black-box model to improve interpretability. In theory, this improved interpretability should allow for the model to be "read" better and see where any biases may come from. However, this method still reduces accuracy compared to a black-box model (albeit to a lesser degree), and does not solve any issues occurring within the original black-box model. These will be further discussed under the section "Grey-box Models".

4.4 ROT/Dark Data

One of the data governance risks that gets greatly exacerbated through the use of Advanced Analytics is that of ROT (Redundant, Obsolete and Trivial) files [5]. Most organisations deal with such files to some degree - consider the average Sharepoint or Drive, filled with duplicate files and unused old projects - but projects using Advanced Analytics should be especially cautious of them. Not only do the large swaths of unstructured and undocumented data in many AA projects mean that ROT files are more likely to creep in, but as outcomes are based on the input data, putting ROT files into Advanced Analytics models tends to give you redundant, obsolete and trivial results.

Closely related to this idea of ROT files is that of dark data. This is data that is acquired and stored within an organisation, but never used for processing. Most organisations are affected by this; 66% of IT managers report that over half their data is considered dark data, with rates only increasing over time [20]. IBM estimates that up to 90% of all sensor data from IoT devices can be considered dark data [23]¿. There are a variety of reasons as to why companies are incapable of dealing with this data; either they lack the proper tools to unlock value from this data, there is too much of it to properly analyse, or the data is simply incomplete.

Since many concepts surrounding Advanced Analytics require data to be processed for decisions in close to real-time, allowing data that can quickly become valueless and obsolete to exist within the organisation can be dangerous. Additionally, since much of this data is often unstructured, there is the risk of sensitive data ending up as dark data, which can be dangerous in the case of data breaches (this is all in the addition to the wasted energy of data centres that store this data [44]). Apart from perilous situations, there are also opportunities; all of this data that is currently not being processed, may

in the future be used by an Advanced Analytics model. When properly collected and governed, this can lead to improved processing results.

Apart from wasted energy, there are also the costs of running the data centres that store this data. Although data storage is generally believed to be cheap, the reality differs; the actual cost is expected to increase year-over-year, even with falling hardware costs [45]. This mistaken notion of cheap storage is due to focussing on the hardware costs of storing data, but ignoring many other aspects of data storage that provide additional costs. These include costs such as maintenance, storage, outages, energy and software, which are all expected to rise over time [45]. These costs are expressed in the Total Cost of Ownership, which is estimated to be five to seven times higher than the hardware acquisition costs [46].

Luckily, these costs can also be turned into an opportunity, as being able to avoid these costs not only saves money, but also provides a clear, quantifiable argument to upper management for setting up a data governance program (this is discussed in depth in the section "Performance Management").

4.5 The Role of the Data Steward

When discussing the aforementioned overlapping of fields integral to data governance, there are few roles that capture this overlap as well as the role of the data steward. Data stewards tend to live on the line of business and IT, "speaking the language of IT and translating it back to the business" [6]. The role has been described as requiring "the patience of a kindergarten teacher and the ability to successfully negotiate a hostage situation." [47].

As discussed in the section "Introduction to Data Governance" The data stewardship role is mostly defined by holding the hands-on responsibility for managing the data. This is done in conjunction with the data owner, who is responsible for the proper handling and validation of access and alterations to this data[11]. Normally, the role of data steward is already quite complex, as it requires a simultaneous understanding of the technical aspects of the data processing, while also being competent in translating this information to business personnel that lack this technical knowledge. Generally, data stewards are invited or appointed from within the organisation, because they are likely to have a proper understanding of the dataset which they are to steward [14, p. 142]. Normally this is a good idea, as these people already take care of their data in their normal work, and are thus likely to show knowledge and responsibility over this specific dataset. It also means that for classical, well designed and properly structured datasets, they will have a good understanding of in what way it should and shouldn't be processed.

However, this is also where classical data governance may fall short of the demands created by Advanced Analytics. Due to the complex technical nature of most AA models, the demands on the technical knowledge of the data stewards are raised even higher. When running an algorithm on enormous amounts of data from different departments, the dataset leaves the scope at which human beings can still properly manage and understand the dataset they are discussing. Especially when the structure in the data is

[11] Data owners are sometimes supported by a team of data stewards with varying backgrounds, ranging from a focus on the business to the IT perspective (and everything in between). This multi-steward structure may become more prominent in future iterations of data governance models with AA-implementations. However, the risk here is that the knowledge gap between data stewards becomes too big, requiring an extra role to function as a link between the different data stewards.

lacking and there is a high growth rate on the data (and thus a high processing speed), things are likely to fall outside of the knowledge scope of a data steward.

All of this hints at the fact that the classical structure of stewardship may no longer be fitting when combined with Advanced Analytics, and that it may be wiser to focus on owner- and stewardship of the algorithms that process this data, rather than the large, fast-moving dataset itself (see section "Analytics Governance").

4.6 The Role of the Chief Data Officer, Chief Information Officer and Chief Technology Officer

The role of the Chief Data Officer – commonly referred to as CDO[12] – is a rather young role in most organisations. The first CDO wasn't appointed until 2002 for Capital One, and recent surveys show that only 21% of the top 2500 publicly traded companies have a CDO, half of them appointed since 2019 [48]. The CDO is responsible for overseeing a variety of tasks such as data management, data governance, and creating data strategies. For most CDO's, the value creation aspect to which these tasks are aimed mostly revolved around Business Intelligence. However, the growing importance of big data and AA are also influencing the requirements for a CDO. For example, to create effective data strategies for AA, one must understand the specific characteristics that AA models have, and how to employ them most efficiently within a data strategy. The fact the role of CDO is often combined with that of the Chief Analytics Officer only reiterates this.

Naturally, the position of CDO is relevant for any data governance efforts in an organisation. The CDO is likely to be an important sponsor for setting up a data governance project, and will need to be involved in the conceptualization and execution of these programs. More importantly for this thesis, including Advanced Analytics is likely to increase the importance of the CDO. With more expansive data governance efforts, the importance of a strong sponsor rises, making the CDO even more critical for the data governance effort. Simultaneously, the increased complexity of AA models compared to BI increases the risks of a CDO not fully understanding the data governance efforts, putting the position of sponsor at risk. This is troublesome, as a good business sponsor is key to such critical efforts [5, p. 61].

A role closely related to that of CDO is that of the Chief Information Officer – generally referred to as the CIO. Despite appearing like a role that overlaps with that of the CDO, their responsibilities are quite distinct. The CIO focuses on the creation of new IT tools and technologies, designing IT strategies and IT service management [49]. The CIO ensures that the IT function of the organisations supports the goals of the overall organisation. In essence, it is an alignment role between IT and business goals.

As the odds are that any data governance effort implements some sort of IT tool to assist with the data governance, this role is also critical for any data governance effort. In terms of positioning between IT and business, the role also lines up with data governance conceptually. As the CDO role is a relatively new role, the CIO role is more likely to be in a trusted and experienced position within the company. This can help in setting up any data governance effort, as this trust can be crucial for communicating with higher-

[12]Not to be confused with the role of Chief Digital Officer. While the Chief Data Officer focuses on what the organisation captures, retains and exploits, the Chief Digital Officer focuses on the digital transformations and the digital strategy itself.

ups such as the CEO, or communicating with employees when there is pushback for the Data governance effort (see "Cultural Resistance"). Although all of this already holds for classical data governance efforts, the fact that data governance efforts for AA are a bit newer and more complex may increase the impact and importance of this trust-relationship. This collaboration between the CDO and CIO can make or break data governance efforts that might feel invasive to those interacting with it at first.

Lastly, we have the role of the Chief Technology Officer, commonly referred to as the CTO. This function is mostly focussed on the technological side of the organisation - including Advanced Analytics. The CTO uses the organisations' capital to invest in technologies that can serve the organisation. For our subject, this means the CTO can have an influence on developing AA techniques that the organisation uses. Naturally, decisions made in this process will also influence the role of data governance; if investments are made in powerful yet non-transparent technologies, this might have a negative effect on any data governance efforts.

As one can see, all these roles appear to be overlapping at times, yet all have their distinct influences on how technologies are handled within an organisation, and thus all influence the role of data governance in the organisations. With how Advanced Analytics is developing, any issues within these respective roles - or communications between them - can be disastrous for a data governance effort.

4.7 Scope

One roadblock that often occurs in setting up a proper data governance program is setting the right scope for the endeavour [5, p. 43]. Organisations that are unaware of how deeply data permeates their organisation have the tendency to treat data governance as an IT project [50], localised within a separate department that lacks interaction with divisions such as business and HR. Unfortunately, this is also where many efforts to incorporate data governance in an organisation fail. As data pervades most aspects and layers of an organisation, so should data governance.

As with many elements in this thesis, scope is a factor where the inclusion of Advanced Analytics basically requires data governance on steroids; with how many parts of an organisation these models can require information from - and with any issues hurting the accuracy and reliability of the models - every decision about the scope needs to be thoroughly considered. For example, if issues occur in the data creation stage (for example, a bank creating client data) and this part of the company is not included in the scope of data governance, issues in the outcomes of the AA model might be hard to pin down (or worse, remain unknown to the users of the models). Additionally, when dealing with sensitive data that might lead to fines when improperly handled, regulatory and legal division will also need to be involved in the scope of the data governance effort.

4.8 Cultural Resistance

One thing literature surrounding data governance tends to mention a lot is a change of culture [29, 51, 52]. Although it may seem that this subject has little to do with theoretical research in information sciences, it is important to address both the human- and business aspects in an organisation. And as almost all of data governance is implicitly practical (nobody does data governance for a non-existing business), it is important to

address the risk of creating a paper tiger. No matter how good your data governance ideas are, if the employees of the organisation dismiss them because they do not understand them or would rather work around them, they will end up unused and forgotten within a short period of time. Part of working in organisations entails working with the people in the organisation, instead of blindly focussing on organisational design and the efficiency of algorithms.

A large part of this issue lies in the (seemingly) added costs from data governance efforts; the idea seems interesting when it is still a fresh new concept, but when the reality of investments and time costs of stewardship sets in, many organisations become hesitant about the programs, and try to dilute them until they no longer serve their original purpose. Most people tend to already be hesitant of change, especially if this includes (the appearance of) more work in the form of data governance roles[13]. Since data governance itself does not create a direct income stream, it can be hard to properly address the importance of it when the implementation starts (this thesis elaborates on this under "Performance Management"). And since data governance requires a long-term, company-wide approach to truly succeed, any part of the data chain failing to meet the necessary standards can bring down the entire data governance effort with them.

Normally, this is already a challenge, and adding Advanced Analytics in the mix only increases the problem. Not only do the large amounts of unstructured data from all parts of the organisation require a culture of people caring for proper data governance, it also adds the challenge of the Advanced Analytics team itself.

As we'll see later under the section "Model Selection", part of data governance for Advanced Analytics will require the consideration of the correct AA model with enough interpretability, even if this reduces raw model performance. As such, you are not only asking them for extra governance or an extra function, but they must also be capable of making this consideration between performance and usability. However, these departments tend to be highly focused on performance[14], making the culture of such a team especially important. If the team does not have a proper understanding of why this endeavour is important, it can eventually become a problem for the data governance effort over the entire organisation. As mentioned, the nature of data governance efforts allows it to wither away by teams not properly managing their data, or in the case of Advanced Analytics, improper model selections.

[13] note that for most organisations, an explicit governance role such as stewardship is not actually more work, but rather a formalisation of a role that already existed informally.

[14] One thing that came out of the literature study done for this thesis is that very little work is done in the interpretability of AA systems. Most research focused on the raw performance side of these models, ignoring factors like useability and interpretability. This seems symptomatic of a general theme within this field, where anything that does not seem to result in performance one way or another is waived to the side.

5 Possible Solutions

5.1 Centres of Excellence

One important aspect of improving the incorporation of data governance in an organisation is a Centre of Excellence (generally referred to as a CoE). A Centre of excellence is a group of experts who communicate and collaborate to support activities throughout the organisation. [53, 54]. For Advanced Analytics, this activity is likely to be the use-cases for AA within the organisation, who can now treat the specific task of data governance on Advanced Analytics as a core activity instead of a side activity.

There are multiple advantages to this. Because communication is now centralised, it becomes easier for different business functions to understand one another. Especially in segmented organisations where systems overlap, barriers for progress can quickly appear when communication isn't optimal. This centralised communication can also help in overall collaborations between different departments and business functions. This extra communication may also lead to more unified and consistent structures and tools across the organisation, ensuring everyone in the organisation is one the same track. Especially in the field of data governance and AA, there are likely to be a lot of different branches involved, where communication about the data governance tools – and feedback on how the different departments implement them - is highly valuable.

As mentioned before, data governance can also struggle with proper upkeep over a longer period of time; if data governance takes a lot of effort and employees do not understand its benefits, it is likely to slowly die out. CoEs also provide the opportunity to consistently meet and discuss goals and opportunities to better achieve those goals.

Conceptually, it may seem that the idea of explicit and centralised teams clashes with the idea of having invisible data governance interwoven with the functioning of the organisation. However, the aim is to have a centralised yet diverse (in terms of function) group of people, that can assist in collaborating across the entire organisation. This centralised communication can help accelerate the path to this "invisible data governance". Most organisations have more than one CoE, possibly even one for every use-case. This could be a consideration for data governance on Advanced Analytics, where a specific Centre of Excellence can be created for just this use-case. This CoE could contain data analysts and architects from both data management and governance branches, and data engineers from the AA fields [55] (and many more).

5.2 Analytics Governance

In traditional data governance setups, the factor that is governed is data. Although this comment may seem superfluous, it does raise an important question for Advanced Analytics. With the amount of incoming data and the pace at which this data renews itself, attempting to only govern the data itself may not be the wisest approach to ensuring proper analytics outcomes. Instead, one may focus on governing the analytics processing models themselves [56]. This approach is generally known as analytics governance [57].

One example of classical data governance not covering enough ground in terms of algorithm usage is that of the Target pregnancy scandal [58]. In this case, the pregnancy of a teenage girl was revealed to her father, before she was given the opportunity to inform

the people around her of this pregnancy. This was possible due to Target's targeted promotional discounts; their analytics system was capable of inferring a possible pregnancy based on previous purchases on the account, and presented personalised discounts for baby items.

Despite the obvious negative implications for the organisation, no actual issues over data governance were present in this incident. The problem was not caused by incomplete or incorrect information, outdated data flows or a lacking policy scope. In reality, this data processing policy is in part considered responsible for their increased revenue in the baby segment from Target over the last decade [59]. Rather, the issue lies in the model itself, and how it implements its algorithm in conjunction with the data, IT, and business processes.

The field of analytics governance is still very fresh, with data analytics literature almost exclusively referring to governance of data. Whether an approach towards analytics governance will have a framework comparable to that of classical data governance remains to be seen, but some concepts can certainly be translated. Especially the data owner-user-steward structure discussed in the introduction is well suited for transforming into an analytics governance structure. In a larger context, the analytics governance may form a data pipeline with the data governance framework, where data access requests that are approved in the data governance framework setup are fed to this analytics governance framework, where specific data usage within a specific model also requires access approval.

As mentioned in the section "Introduction to Data Governance Roles", this focus on ownership of algorithms neatly fits the structure of a RACI-matrix, where the "responsible" and "accountable" roles align nicely with the governance roles of "owner" and "user". Although classical data governance may focus more on the roles and responsibilities surrounding the data itself, applying data governance towards advanced analytics may require a similar approach towards model governance. This can be done by including tasks surrounding the governance of models into these RACI-matrices.

5.3 Data Quality Management Policies

As mentioned, Advanced Analytics – despite its immense potential and power – is still little more than transforming inputs (in our case, business data) into a more valuable output. Whether the issue is duplicate data, ROT/dark data or inherently biased data, high quality data is paramount to the usability of these systems; after all, no level of governance and management is going to improve inherently bad data. This is also where one of the main challenges in Advanced Analytics lies; with large amounts of data coming from unstructured sources, it can be hard to get this data properly sanitised and usable into your systems.

One paper that discussed some data quality management policies is the case analysis on the National Pension Service of South Korea [60]¿. As the South Korean government aims to disclose data to private enterprises to support new commercial services, big data solutions are likely to play a role in these services. However, previous experiences with big data led to problems with security, accuracy, and unreasonable expectations. Thus, there was a need to devise strategies that allow big data-related solutions using this data. Among other things, this research refers to [60, p. 388] four separate data attributes relevant for data quality in big data;

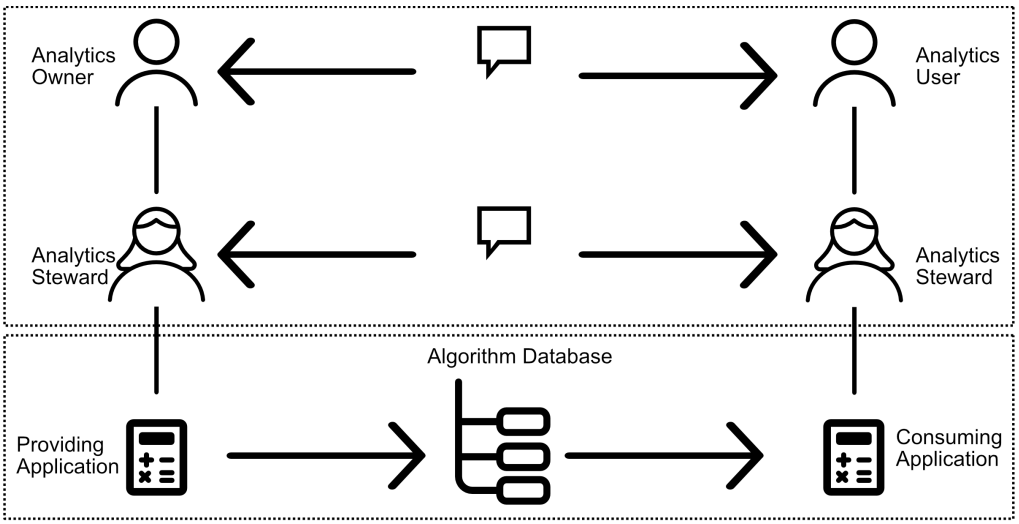

Figure 4: A Preliminary Analytics Governance framework, adjusted from Bas van Gils' [14] data governance framework. As data storage is not a component of analytics governance, the bottom section now represent the selected analytics algorithm from a list of possible algorithms

- Timeliness: this refers to the timing of the data being prepared and processed, so that it can still create value. If the timing is off (i.e. no longer suited for Advanced Analytics), it not only loses its value, it can also distort the outcomes of Advanced Analytics models. This makes it hard to safely use the models, as one is never sure if the outputs are based on outdated information.

In terms of policies, setting proper standards towards the creation and verification of the timeliness of data can ensure that only data that has a high enough timeliness quality standard is used in processing. Preferably, this is a largely automated process, as the timescale for Advanced Analytics may be too short for manual human intervention (especially when dealing with something like live sensor data).

- Reliability: this refers to how much the data can be trusted, and whether it can *"prove the validity of analysis results"* [60].

Although timeliness is already a part of the reliability of the data, this attribute takes a slightly broader approach. This can make it a bit harder to properly define policies for this attribute; after all, how can one enforce validity if there is no proper definition for "valid" (in contrast to the possibility of timestamps for timeliness)? How this is to be approached is highly dependent on the context of the organisation, most importantly what data they are processing and for what purpose.

- Meaningfulness; Meaningfulness refers to whether the data can *"provide meaning as a topic appropriate to the purpose of analysis"*. Once again, the risk is not just that the organisation's resources are being wasted on the storage, maintenance and advanced processing on meaningless data, but that it is also not clear whether the outcome is meaningful or not, hindering the possibility to make informed decisions.

As a policy for this attribute, it is closely linked to the issue of ROT/dark data. Data that builds up without a function is best handled when entering the system by some filtering mechanism, instead of being removed later on.

- Sufficiency; lastly, sufficiency is a quality measure of quantity; is there enough data that the purpose of the analysis (and hence the organisation) can be achieved?

Although data quantity may seem easier to define for policies, the attribute might become a bit complex. Quantity is not just measured in file-sizes or the amount of files, but also the depth and value of the contents of the files. This may require some data transformation prior to data quality controls to ensure that the definition of "enough data" is not abused.

5.4 Storage Tier System

With the scale and unstructured format of these big data projects, it can be hard to properly figure out what data carries importance towards the organisation in terms of security hazards. A tier system of data value can help not only in reducing cost (as less sensitive data can be stored without the costs of added protection), but also prioritising what data requires extra forms of protection and maintenance. Although this leans more towards data management than governance, prioritisation of data can help ensure every data asset receives a proper level of auditing.

However, such a tier system with the scale and unstructured format of big data projects can't be maintained manually, as this would eat into limited resources from the organisation. Instead, this approach requires some sort of filtering system, which checks the sensitivity of the data and where (or whether) it should be stored. What the parameters of such a filter consist of is largely dependent on the organisation itself and how they intend to use the data, but the aforementioned "Data Quality Management Policies" can form a guideline in this. For example, "timeliness" could be used as a factor for incoming data, which compares timestamps of data entry to some threshold to see how timely it is. It could also be done in the context of the existing data lake (or mesh). For example, "sufficiency" can be measured through whether it adds sufficiency to the existing data set, or whether already existing data makes the new data redundant in the context of the data set.

5.5 Data Lakes and Data Mesh Architectures

Structurally, there are many ways to set up the hierarchy of data governance and ownership, with the solutions laying on the spectrum of centralised to decentralised structures. In a centralised setup, all power (or in this case, data storage and management) is handled from a single location, maintaining authority on data from a top-down position. Any employee that wants to access data will have to move through this database.

Data lakes are a good example of a centralised structure, which is meant to handle large amounts of data with varying structures. These are generally built with the use-cases of big data and Advanced Analytics in mind, allowing data to be stored without requiring structure transformations first, and sometimes allowing data analysis directly from the lake itself.

In a more decentralised setup, the data structure is instead spread over multiple databases over different locations. The communication is then generally done through some centralised location, such as a network or some dashboard. Here, the governance itself is led by leaders of departments, or through practitioners themselves. Some have expressed the idea of considering this a "data community", as data users also take responsibility for their assets so they can be properly reused by others. One big advantage here is that the decentralised structure removes many layers of bureaucracy to upper layers of governance, as everyone can independently access data and gain insights from it.

One of the newer technologies for decentralised data architectures is that of data mesh. These were born out of the limitations of data lakes; for large organisations dealing with big data, making data products out of data lakes while adhering to company standards becomes too time-consuming [61]. This is a common problem with centralised solutions of any kind; when size and complexity of a governed asset grows, centralised solutions cannot cope with the increased needs, and a backlog of requests builds up. Centralised teams are also generally not aware of all domain knowledge, possibly reducing the quality of the product [61] .This is where the power of decentralised solution lies; as most domains know their own assets best, it makes sense that they are also responsible for managing this asset.

In a data mesh, organisation domains are allowed to manage their own data, and are then backed by a *"central and self-service data infrastructure"* [61]. The idea is that built on top of data warehouses (an older, less flexible form of a data lake), are different pipelines that are managed by owners of that specific domain. This domain-oriented design is the key to the power of a data mesh network; it gives data owners the autonomy over their domain, while reducing pressure on centralised pipelines [62]. Apart from the competitive advantage this provides, it also helps in setting up data governance efforts. When teams are responsible for the data they manage, they generally have a better feel for their data governance needs than a centralised team may have.

Despite the notion that decentralised solutions are generally the better solution, they are not a one-size-fits-all approach to data governance and architecture. Smaller organisations with few (or no) separate divisions in particular may thrive under the simplified structure of a centralised approach [45,63]. It also reduces data redundancy (as everyone is working from a single database).

5.6 Federated Data Governance

A solution in the middle of the centralised-decentralised spectrum is that of federated data governance; here, some elements of the organisation are handled centrally top-down, yet other functions are handled at an operational level. For data governance, this generally means that some semblance of governance is handled top down (such as standards and policies), yet more local teams are given the power and autonomy to implement things based on their particular needs and resources [5, p. 25]. The advantage is that despite having consistency across the organisation, separate departments are still able to optimise for their respective environment, improving efficiency and performance. Especially if an organisation has highly varying needs of data governance between departments, federated data governance can be very helpful. It allows departments handling more tightly regulated data to implement strict policies and prioritise tight governance, while allowing other departments to govern more loosely, improving performance.

The level of federation can also be based on the sensitivity of the data (or other forms of risk assessment), where the scaling of sensitivity then determines how much top-down influence is necessary for proper data governance - similar to how the storage tier system works. These two approaches can then be combined into a single tier system, entailing both storage resource investments and top-down governance influences.

The decision to use federated data governance is one that needs to be well-considered before being implemented: Although federation can help in preventing control issues and maintaining the scale of governance, it can also lead to issues in hierarchy; due to the freedoms provided by the federated section of the organisation, they may clash with other federated section with which one cooperates, or even the highest centralised body. Other systems (such as the US) use an issue-resolving body to deal with these discrepancies (such as the US Congress and house of representatives), but that doesn't remove the risk of clashes in terms of execution and implementation of policies between lower-and higher level programs.

5.7 Automated Governance

A big part of proper data governance is not only the ensuring of quality, but also the speed at which the auditing can be done. A highly accurate auditing system is still useless if execution takes so long that large backlogs of requests build up. Especially when working with manual approval systems with large amounts of requests, projects that require specific pieces of data can grind to a halt (not to mention the fact that these approvals are often done by employees with full-time jobs alongside their stewarding function, slowing down those roles as well). If parts of the auditing system can be automated to pick the low hanging fruit in terms of approval,the focus is now allowed to shift to approving manual cases faster and better.

Although not every form of data request is open to such automation – some require case-by-case reviews by law, and organisations may want extra governance for some cases (see "Federated Data Governance") – being able to pre-filter some simple data access requests can reduce workloads on the employees handling these requests, while simultaneously ensuring that projects can keep moving forward. Even automating part of the task (for example, creating a grade based on the requests characteristics so stewards can at a glance see whether a data access request is reasonable) can improve the response time on data access requests.

5.8 Model Selection

Before discussing how to handle the data governance of challenging concepts such as black-box models, it may be wise to first ask the question why one is using these models in the first place. Despite our best efforts to ensure proper data quality, there is always the risk of data ending up in unassigned positions. Especially when the results of this can be considered severe[15], it may be wise to consider the enforcement of more interpretable models.

[15] In Risk Assessment processes, this is generally measured through a Risk Assessment Matrix, where risk is defined through the combination of the probability of a risk and the impact of a risk [64].

Those who have little experience with the usage of these models may be tempted to simply select the solution with the highest theoretical performance, but there are other factors to consider. A lack of interpretability may actually be more costly than the lost performance, especially if one is dealing with sensitive data, or when the performance gap between black-box models and white-box models is rather small. This is a decision that the legal and forensics world has dealt with numerous times in the past [65] [66], where white-box based models were good enough to find results, without dealing with the interpretability issues of black-box models [67]. Additionally, increased interpretability can allow for more communication with a model, improving [68]. If this added communication and interpretability can help find flaws in how a model reaches its conclusion, using an alternative model instead of blindly trusting black-box models can help improve performance and trust in the model.

Note: This is something that technologies like Advanced Analytics and Artificial Intelligence struggle with in general; upper management being distracted by the shiniest new technology, without any understanding of where and why they should be used. For example, in Japan, 49% of organisations believe edge computing will be vital for their organisation, yet 58% do not have a use-case in development. This is not inherently a problem – many technologies are built before they find the problem they will solve – but it is something to keep in mind when considering why an organisation uses complex analytics models; does their use-case truly require it, or were they simply enticed by new technologies?

Although proper model selection does not solve all the issues and tends to circumvent the problem, minimization of risk and impact can make the execution of data governance a lot easier, and reduce the impact of when data governance does go wrong.

5.9 Grey-Box Models

One extension made towards models that can help in the interpretability and governance is that of grey-box models [43, 69]. These are models that aim to combine the accuracy and complexity of black-box models with the interpretability of white-box models. In contrast to a black-box model, a white-box model retains the interpretability that is lost in black-box models, as one can interpret it simply by looking at the internal parameters. An example of a white-box model is a decision tree[16], in which one can simply walk through the tree to see why elements are categorised in a certain way. On their own, decision trees (or rule-based systems) are also not perfect, as their trees (or rules) can quickly become complex and too large to easily understand. However, the combination of the two model types aims to get the best of both worlds.

Grey-box models tend to work by starting out with a black-box model, which outputs labelled data on which the white-box system is then trained [43, p. 6], [69]. Unfortunately, these systems are not suitable for every use-case; they work best with limited labelled data and lots of unlabeled data, and are less accurate than black-box models (especially when the initial dataset is too small). Nevertheless, future projects should consider the usage of these models, especially when interpretability is key and a small accuracy loss is acceptable. This added transparency and interpretability greatly eases the load on the data governance effort.

[16] Other examples are linear regression models, Bayesian networks and fuzzy cognitive maps [43, p. 3].

Once again, one can expand upon these models by expanding the decision tree with fuzzy logic [43, p. 9]. Normally, decision trees work on the basis of either numeric values (e.g. Weight ¿ 60kg), or nominal distinctions. These are non-ordered categories of a variable (for example, Gender = Male). Fuzzy logic transforms this into a balance of probabilities towards a state, where something is now "true with a probability of x%", instead of one branch being completely true. Combining this with linguistic hedges (e.g. something is "somewhat true") can provide even better interpretability for those with less knowledge of these models.

5.10 Accountability in Usage: Teaching Advanced Analytics to Business Users

One of the things that seems especially important for data governance around Advanced Analytics is accountability of model usage. As data governance is primarily about how people handle data - not the data itself-, part of this governance means that people need to be accountable in how they use these systems and their outcomes. One of the risks that come with new and shiny technologies like Advanced Analytics is that the users -especially those uninformed about its inner workings - start to see it less as a tool and more as a perfect prediction of the future. As powerful and accurate as these techniques may be, at the end they do little more than transform data into other data, not create some truthful image of how the world works. Looking beyond the hype of these systems and knowing how these ideas come to light is especially important. Part of a governance strategy may be to teach non-IT departments what these systems fundamentally do, so that the decision involving these systems can be made more responsibly. This does not mean that every person in HR needs to have a deep understanding of how to build neural networks, just what these systems are actually doing at surface level and how that may affect the decision you make with them (the section "The Black-box Problem" can be of use in this).

A more concrete way to measure this proper usage may be to quantify the AA model outcomes for the users. Translating important characteristics of these outcomes such as the data quality and explainability into a single score can be a helpful tool for the user, so they can at a glance understand how heavily they can rely on the outcome of the model.

5.11 Aligning Business and IT

As mentioned before, data governance – although technically inclined – is at its core a business process, not an IT process. It focuses on improving business processes and achieves this through the use of policies, standards and guidelines [60]. Hence, proper alignment between business and IT is vital in ensuring a consistent approach, implementation and cooperation towards data governance.

However, consistency and centralization are concepts that need to be handled with care. Although they are terms that generally provide positive connotations and work well with upper management, one has to remember that not all departments of an organisation are built equal, and that excessively chasing consistent policies and standards across a wide organisation may hamper the specifics of departments. For example, data retention needs for HR may vary wildly from Finance, as their data can have vastly different life-

cycles. The key here is to find enough consistency across the organisation to maintain easy communication between departments, while leaving room for department-specific implementation to support their respective needs.

These excessive data governance policies also present another risk. Most policies tend to be built with specific use-cases in mind, helping proper guidance for the most used processes. But with technologies as young and complex as Advanced Analytics, much of the value-creation can come from using these technologies in unexpected ways to create interesting outputs[17]. If an organisation is too bent on enforcing policies and standards for specific situations, it runs the risk of losing out on these valuable outputs, defeating the purpose of owning this data in the first place. In essence, an organisation should be finding a balance between freedom of interpretation, without leading to lack of continuity.

5.12 Rethinking the Separation between Management and Governance

The introduction of this thesis discussed the separation between governance and management, expressed through the "V" framework. The reason this separation exists in the first place is pretty clear; one of the main roles of any audit is to provide an assurance towards the auditee itself, in the form of a competent, impartial judgement [72]. If the auditee itself is allowed to influence the auditing process, the quality and integrity of the process is diminished. In classical auditing processes such as a financial audit, this separation inherently works well, because the separation between the two processes already exists.

However, this separation may not be as clear for Advanced Analytics, especially when moving from a theoretical framework to real-life usage. Even in this thesis, the border between governance and management is at times muddy. For example, the aforementioned data quality management policies come from the world of data governance, yet strongly influence how data itself is handled in practice (and thus also occupies the realm of data management). Maintaining stronger feedback loops and a tighter relationship between management and governance can vastly improve both processes. After all, if governance can help shape how data is managed, it also reduces the load on the data governance effort, which in turn can assist with making data governance an invisible part of the organisation.

Of course, one should be apprehensive about losing the separation between governance and management, as the impartiality of governance is still a key factor in improving data management. Perhaps a balance between both forms is a more attractive option: A small team of both data management and data governance experts -perhaps a CoE - who work closely with both the governance and management teams, and incorporate elements of the management feedback into the governance setup.

[17] The EU seems to agree on this, as their Data Governance Act from 2020 [70] aims to increase the amount of data available for re-use within the EU, by *"allowing public sector data to be used for purposes different than the ones for which the data was originally collected"*. The Data Act from 2022 [71] builds on this through further clarifications.

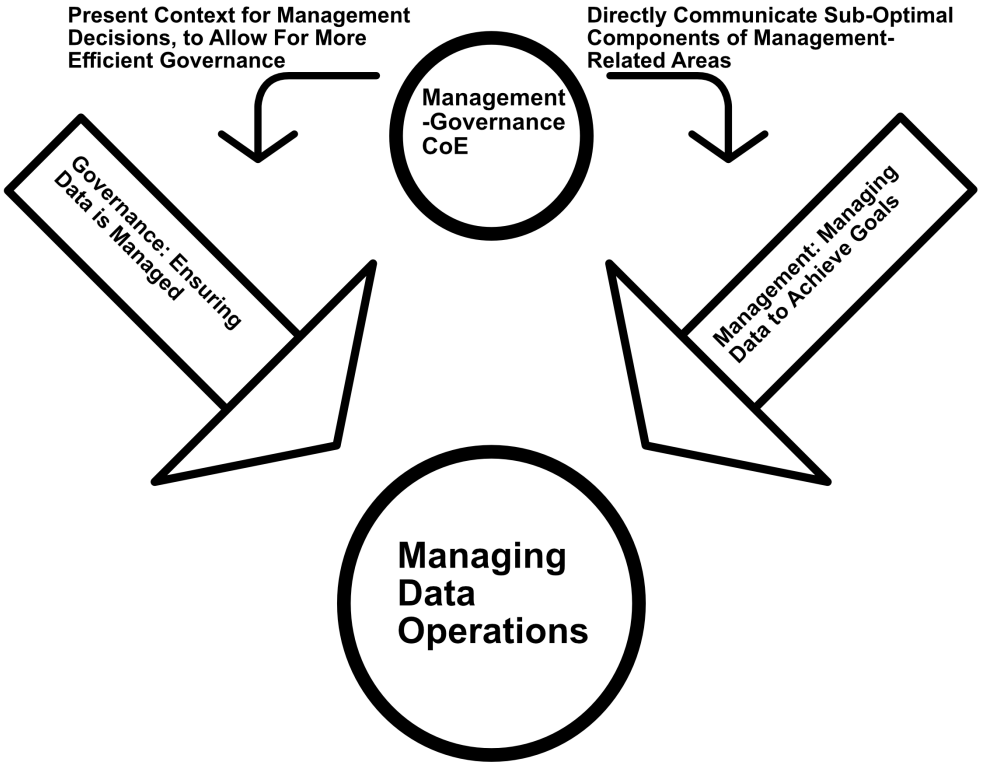

Figure 5: An adjusted version of "the V", with a central CoE providing towards both data management and data governance departments

5.13 Managing Culture

As mentioned in the section "Cultural Resistance", it is important to include the human (and thus cultural) aspect of an organisation when discussing a data governance effort. Part of this effort should be to ensure that everyone involved understands the importance of data governance, and why it might require an upfront investment of resources. This includes convincing upper management of the value of proper data governance and ownership of data. One possible perspective here is to use a metric every business owner understands; cost.

Clarifying the actual advantages in financial terms not only helps in convincing the organisation to start up a proper data governance effort Advanced Analytics, it also helps in proper maintenance of this effort. When someone fully understands the cost of managing data and is held responsible for this management, it reduces the risk of improper data governance. This is also where companies start to see the advantages of data governance become more well defined, as companies that do data governance spend less time reacting to data-related issues than those who don't, which in turn can be spent on running the organisation.

5.13.1 Performance Management

To make a strong case for data governance on Advanced Analytics within a company and to ensure that it is actually working, some form of performance tracking and management is very useful. Especially in a corporate culture that requires a quantified showing of results to determine the value of a program, being able to present some metric of performance for data governance can be a very convincing argument for upper management and employees having to implement these policies [29]. Unfortunately, results from data governance efforts are a bit harder to define than results from sales departments, or stock dividends. However, here are some possible quantifications:

- Reduced losses from lawsuits and other liabilities: If improper data governance leads to a situation in which an organisation can be held liable for the costs of an accident (e.g. improper data usage), the investment in data governance can be seen as the leading cause for avoiding these costs. This can range from privacy violations, to civil and regulatory liabilities due to poor management.
- Increased productivity for time spent on value creation: As mentioned, data engineers spend a majority of their time not on creating value from data, but on finding the correct data and ensuring it is ready for usage. In essence, this is time wasted by highly paid employees on company time. If data governance is executed properly, the increased productivity from opening time up for value creation from data can be considered a performance increase through data governance.
- Overall increased productivity: The value creation productivity mentioned here does not exist in isolation; as data processing is a chain, it creates a ripple effect all throughout the organisation (and even any possible third parties). Due to the time saved, information can be delivered faster, allowing for quicker decisions by management. The time saved can also be filled with more time-consuming data processing methods that might cost more time, but may also yield better results. In theory, all of these advantages can be attributed to data governance.

- Reduced storage and maintenance costs: As mentioned in the section "ROT/Dark Data", the total cost of ownership is five to seven times higher than hardware acquisition costs [46]. This is a nice number, as if you have any indication of how much ROT/Dark data data governance has kept out of the organisation, you can use the TCO figure to get a pretty good approximation of reduced costs due to data governance.
- Improved decision making: with data governance hopefully leading to better information to drive decision making, it can prevent losses (or missed income) that would've occurred without this better information. For example, if poor decision making would have led to poor inventory management (e.g. expecting too few or little sales for the existing inventory), an argument can be made that data governance directly led to the avoidance of these reduced profits.
- Reduced losses from bad PR: When news breaks out that an organisation has poorly handled information, this almost certainly hurts their customer's confidence in their competence - and in turn, their bottom line. If data governance can work as a preventative tool against this, then any costs that this negative PR would have provided can be considered in quantifying the data governance's value.

Admittedly, the added value of some quantifications can be hard to properly define (after all, one cannot accurately define the cost of a theoretical liability, especially without knowing how likely it is that it would've occurred). However, being able to provide some semblance of quantified value and performance tracking can be enough to make upper management understand the value of these programs (especially the risk aversion factors can be powerful for large, risk-averse organisations). Additionally, many of the measures are also relevant to classical data governance. With Advanced Analytics building on many of the concepts of classical BI, its data governance is often going to see similar measures and results, with the demands on them turned up to 11 due to the characteristics of Advanced Analytics.

Finally, apart from quantifying the value of data governance itself, another important quantification is to properly value the data itself within the organisation [5,45] It is easier to create a culture of accountability and understanding when one can quantify the value of data in some way or another. This is especially true for the starting phases of data governance, where other projects with more quantified goals and resources (finances, time, effort) will be prioritised over the data governance efforts if the data is not properly valued.

5.14 Main Artifact: Governance Framework

As a final artifact, this thesis presents a governance framework, aimed at adapting data governance efforts for Advanced Analytics. As this is a generalised framework, real-life relations are likely to be much more complex and bidirectional, with organisation variables also affecting one another. For example, a better culture surrounding data governance is also likely to improve communications about this subject. However, this model gives a good approximation of what elements influence what components within a general data governance setup.

	Data Quality	Storage Quality	Culture	Communications	Model Usage	Request Response Time	Domain autonomy
Centres of Excellence			X	X			X
Analytics Governance					X		
Data Quality Management Policies	X						
Federated Data Storage		X					X
Data Structure Selection		X	X				X
Automated Governance			X			X	X
Model Selection					X		
Accountable Model Usage			X		X		
Business-IT Alignment							X
Management-Governance CoE		X	X	X			
Performance Management			X	X			

Final Artifact: A conceptual model for data governance in the context of Advanced Analytics, based on the suggested solutions within this thesis. For readability, an additional diagram-based version of this model is added as external file

The main aim of this framework is to present a simple overview of factors that may influence organisational element relevant for the execution of data governance. A short explanation of these organisational elements:

- Data Quality: This represents the quality of the data itself. This includes the existence of ROT/dark data, and any factors discussed under "Data Quality Management Policies".
- Storage Quality: This represents the appropriateness of the data storage structure for the respective task and organisation. This includes the proper selection for centralised vs decentralised structures, and proper federation of data storage if necessary.
- Culture: This represent the organisational culture surrounding data governance, as discussed in "Managing Culture" and "Cultural Resistance". This includes elements like a willingness to invest in proper upkeep, and taking accountability for proper data governance.
- Communications: This represents how efficiently and transparently the organisation can communicate about its data governance and data access requests.

- Model Usage: This represents the appropriateness of the selected model for the task at hand. This includes whether a correct consideration has been made between the accuracy and interpretability of a model, and proper interpretation of the results provided by the model.
- Request response time: This represents the time it takes for data requests to be processed, such as access requests for specific datasets.
- Domain autonomy: This represent how autonomous separate departments can operate from one another. This includes factors like the freedom to alter policies to improve efficiencies specific for their departments (without losing consistency across departments), or the reduction of time loss due to waiting for access requests from other departments

5.15 Factor Selection

As mentioned, the real-life relations in an organisation are likely to be much more complex and much less binary than presented here. Here are some elaborations on the decisions that may not seem intuitively clear:

- Automated Governance → Culture: If data governance can be partially automated, it'll take up less time from both those who normally have to wait for data access permission, and from those who normally have to approve the data access request. Removing time- and effort investments helps in the mission of making data governance invisible, which is likely to improve the culture surrounding data governance.
- Automated Governance → Domain Autonomy: If departments can use automated access requests when retrieving data from different departments, this reduces their dependence on other departments' personnel, improving autonomy.
- Performance Management → Communications: If members of the organisation are presented with a quantified metric derived from performance management, this also allows for discussion and communications surrounding this quantified metric. This may also improve the communications surrounding the success and right direction for data governance as a whole. For example, someone may find a better metric for capturing the success of data governance, which in turn can steer data governance in a better direction.
- Accountable Model Usage → Culture: Apart from leading to better usage of models, teaching Advanced Analytics to Business Users is also likely to lead to a better culture. If business personnel understands why they are using the Advanced Analytics model and what the consequences of improper usage, they are also more likely to feel responsible for proper usage. If done at scale, this responsibility will spread across all personnel, improving data governance culture.

5.16 Usage

The simplest way to approach using this model would be to first consider what elements are most lacking within your organisation, either through an audit or the results of

performance management. Then, address the most important elements through the respective possible solutions presented for these elements in this table. If the table has been given ecological validity through future use, it could also be used a form of validation; if any factor that may influence organisation elements has been altered within the organisation, one can validate the success of this effort by seeing whether the relevant organisation elements also improve.

6 Interview

6.1 Introduction

The interview with Mr Snijders started off with a short introduction of what the position of Chapterlead Data Analytics entails. Mr Snijders described the position as carrying the responsibility for the professional development in the area of Advanced Analytics, in the widest sense of the word; from data scientists, to data science and software engineers who work on embedding the models in software, which makes them robust, performant and controllable. In addition, he's responsible for personnel development and recruitment, which should ensure that they have the right population for enabling the digitalisation within Alliander.

Mr Snijders explained that as a part of their data science endeavour, Advanced Analytics is used widely within the organisation. His chapter consists of 95 people, who have been using Advanced Analytics for the past 8 years. Advanced Analytics usage entails elements such as forecasting, but also use-cases at operational, tactical and strategic levels. This includes (but is not limited to):

- Predicting network congestion and overcapacity, optimising cable placement, and predicting cable routing between compact stations and residential areas.
- Creating a digital sketch of the network itself, allowing for calculations concerning capacity and real-time malfunction predictions.
- Automated checks for whether a potential new customer can be connected to the network, and predictions for what their waiting time would be.
- Assisting the technical contact centre which handles outages, who use data such as weather information and road traffic to determine outage risks and at what capacity the contact centre needs to be.
- Predicting the risk of excavation damages based on data such as soil type, cable usage and density, contractor selection. This prevents redundant excavation controls, reducing personnel strain.

Additionally, HR and Finance departments are also seeing some development in the use of Advanced Analytics.

In terms of data entering the organisation, they combine sensor data with open domain data, core system data and other registers. Data is prepared in bite-size segments to maintain usability. This also includes sensitive data, which has to be handled strictly in order to comply with the GDPR.

Next up, we discussed the cultural attitude towards data governance within Alliander. He mentioned that it tends to go either way; on the one hand, those who want quick access to data are sometimes bothered by the extra time that data governance can add to a request. However, he also mentioned that more and more people are starting to understand the importance of proper data quality:

"You can have a great model, but if the fuel [data] that you add to this model is of a very low quality, then your outcome is also going to have a very low quality".

To close the introduction, we briefly discussed whether they feel their data governance is up to par for the future growth of Advanced Analytics. Mr Snijders described it as a path that they are currently on, but also a process that is never finished. A lot still needs to be improved, especially the usage of a more systemic approach. However, he also mentioned that a decade ago, people were already saying that they weren't ready for this, and that despite the steps made in these years, the growth of the possibilities with Advanced Analytics naturally requires extra steps from data governance.

6.2 Challenges

For the first possible challenge, we discussed the traditional structure of data user/steward/owner. Despite not completely matching up, Alliander does use the basis of this interaction in some form. One of the big changes that occurred in recent years is the addition of the data office; a business unit responsible for data management and data governance. This unit contains personnel such as data architects, data modellers, data stewards etcetera. This unit centralised a lot of these data interactions. I asked whether this data office can be compared to a Centre of Excellence, which he concurred. He also stated that the awareness and importance of these interactions has improved at the software engineers and others involved over recent years.

The next subject was that of black-box models and their usage. Mr Snijders stated that their underlying mathematical models can be made fully transparent. They do take in account the risk of biasing the data. One point he made is that despite the association of data science with neural networks and deep learning, most of their modelling is closer to network calculations, optimisation algorithms, and even very basic processing such as linear regression. Only three teams in their organisation are working on concepts like AI and deep learning algorithms, thus entailing only a very small portion of their organisation. He also mentioned that even though models may not be perfect, the impact of their failures is generally limited. For example, Alliander uses image recognition on their fuse boxes to determine what tools a mechanic needs to bring for reparations. The impact of failure here is relatively low; it only requires an extra trip for different tools if the model fails. Weighing against the reduced preparation time the model provides - from two hours to fifteen minutes per case - means that it still accounts for a net gain of time for the mechanic. The risks taken here depend on the context; when making changes in the network itself, they minimise the risk as much as possible.

After this, we switched to data quality policies. Mr Snijders mentioned that they have such policies, and even do data quality measurements. These look at the quality of specific data elements such as contract- and asset data. Additionally, they also have rules about the intake of their data. They also have varying degrees of policy strictness; the more important the data, the stricter the controls are. For example, sensor data from the network is likely to have very strict policies.

Next up, we discussed the importance of the CIO and the CDO. According to Mr Snijders, The CIO is responsible for the overall digitalisation of the organisation, while the CDO is responsible for the quality and usage of the data. One transformation he sees in these roles is that many organisation are slowly transforming into a tech company, including Alliander itself:

"Even small sized organisation have a CIO these days. You can see the digitalization becoming more and more important, and companies that haven't properly digitised, you

can see them starting to collapse."

The next concept we discussed was a possible shift from data governance to analytics governance. Mr Snijders believes that data scientists need to take responsibility over the outcomes of the algorithms. New data then gets created from these outcomes, with the responsibility for this data given to the creator of the data. Although this process should be monitored - and personnel should be addressed on any data issues - the responsibility itself should be decentralised, instead of at a single business unit like the data office.

The final question surrounding possible challenges was about what challenges they ran into setting up their data governance efforts, and what the main lessons were that they learned. The main point Mr Snijders took was that the concept of a single truth doesn't really exist and that instead, one should focus on finding a similar view and perspective.

6.3 Solutions

After this, the conversation shifted to possible solutions for data governance problems, starting out with data storage structures. Mr Snijders mentioned in the previous question that in the past, they attempted to create one big data lake, but that this didn't work out, and that they switched to a more decentralised approach through a data mesh. This decision was made because it allows the organisation to place responsibilities for data where it can be carried, instead of at a centralised place:

"Our company employs ... approximately 8000 people, you can't verify for everyone what exactly they're doing, so you have to ensure that the checks occur at the right place, so that people that are close [to the data] know what's going on."

We shortly discussed the possibility of automating data access requests. Mr Snijders mentioned that it fully depends on the data; sources that have been granted access previously can be done relatively quickly, but he doesn't believe in fully automating the process for a completely new data source.

Following this, we looped back on the subject of data governance culture, but in the context of non-technical departments like finance and HR. Mr Snijders remarked that finance departments are generally quite thorough in general, so they see the value of proper data governance. HR departments are mostly involved on the privacy-end of data governance, considering they handle sensitive employee- and applicant data. However, they are generally less thorough in data management elements like easily retrievable file storage. Related to this was the data governance culture of data scientists, with Mr Snijders stating that data scientists are most likely to realise the value of proper data governance; if data isn't in a proper state, the outcomes of the models also lose value. One thing he did see was data scientists trying tricks on models to solve small problems, while these should really be solved structurally. For this reason, Alliander has a data counter [data loket] where employees can report any data issues to be resolved.

Next up, we discussed the separation between data management and data governance. Mr Snijders stated that their interactions differed per team; teams where data quality carries high importance have a close relationship with the data governance effort, whereas other teams require little more interaction than following a piece of policy.

As a final point, we discussed performance quantifications of data governance. Mr Snijders recognized the difficulties with how to measure the added value of data governance,

but also questioned the need to measure and value everything within the organisation. He wonders whether sometimes, the resources required to measure everything might be more costly than simply admitting that an organisation sometimes does something that doesn't provide the right value. Finding a balance here is an important task for the data office. Especially since part of data governance is about prevention rather than curing, it can be hard to quantify a cost that doesn't exist because of proper data governance.

6.4 Closing Thoughts

The final questions were related to how data processing would change over time and how data governance needs to adapt to this. Mr Snijders sees data processing going in the direction of more automation, with more direct interactions between systems and with the outside world. Although this should lead to higher quality processing, it also introduces more dependencies. Removing the human in the middle also removes an opportunity for intermediate quality controls, thus requiring proper automated monitoring.

The final question posed was what risks and opportunities he saw in the future growth of data governance. The main pitfall Mr Snijders foresees is data governance becoming too big and complex, leading to people dropping out on data governance, and policies becoming unused. This translation to the workplace is also where he sees a lot of opportunities; if data is handled properly and definitions are coordinated clearly for all parties involved, you can use automated interfaces much more efficiently.

7 Conclusion

This thesis set out to get a grasp on the future of the field of data governance on Advanced Analytics, by discussing the challenges it may present, and proposing possible solutions to these challenges. The first research question to grasp this goal was:

"What are the challenges currently present within data governance on Advanced Analytics?"

After an extensive literature study, this thesis proposes a wide variety of possible challenges for data governance on Advanced Analytics, summarised here:

- Black-box problem: systems that are opaque about their internal working (or are hard to interpret) can provide issues when aiming to do analytics on these system's internals, or explaining decisions that these systems are involved in.
- Stochastic MDM: The randomness involved in the internal optimisation steps of some Advanced Analytics models may clash with the aim to have a "single version of the truth" through Master Data Management.
- Discrimination: Discriminatory biases in data fed to Advanced Analytics models can be reflected in the outcomes of these models, perpetuating these discriminatory biases.
- ROT/Dark Data: The large and unstructured data flow that big data creates can lead to lots of unprocessed and eventually unusable data, which leads to wasted resources through unnecessary storage, and polluted data that negatively affects the outcomes of models.
- The Role of the Data Steward: The scale of the data that Advanced Analytics is involved in is outside of the scope that a human can reasonably manage. This affects the role of the data steward, as it becomes harder to understand how data should be properly processed.
- The Role of the CDO/CIO/CTO: The increased complexity of Advanced Analytics models can make it hard for these officers to understand and maintain their role in supporting the data governance effort.
- Scope: The amount of departments that are involved and affected by Advanced Analytics force data governance efforts to consider a rather large scope, which means there are more places where issues can occur, thus requiring more coordination.
- Cultural Resistance: The upfront investment of time and costs is likely to scare off organisations setting up a data governance effort, especially if the organisation is already hesitant of change. This can lead to the data governance effort being diluted until it no longer serves its purpose.

Building on this question, the second sub-question stated:

"What possible solutions exist for the challenges currently present within data governance applied to Advanced Analytics?"

Again, the solutions found are summarised here:

- Centres of Excellence: CoE's provide an opportunity to centralise communications and treat data governance as a core activity.
- Analytics Governance: Analytics governance allows one to expand the governance structure, by focussing on governing the models themselves instead of the data.
- Data Quality Management Policies: Using policies to guarantee high quality data can ensure that models are not negatively influenced by low-quality data.
- Storage Tier System: Prioritising your most sensitive and critical data in terms of protection and maintenance can prevent issues with security, while not spending unnecessary resources on superfluous measures for less relevant data.
- Data Lakes and Data Mesh Architectures: selecting the proper data storage structure can guarantee that accountability is kept close to the source, while still allowing teams easy access to their data.
- Federated Data Governance: Providing some level of top-down governance guarantees consistency across the organisation, while separate departments can still optimise their respective environments.
- Automated Governance: automating part of the governance allows organisations to reduce workloads for personnel handling data access requests, while simultaneously reducing time projects spend in limbo waiting for data access approval.
- Model Selection: Taking the effort to select a suitable model for each task can improve interpretability of models, while losing very little accuracy. This process also includes the possible selection of grey-box models, which combine the accuracy of black-box models with the interpretability of white-box models.
- Accountability in Usage: Educating non-technical personnel on how to properly interpret and use the outcomes of Advanced Analytics models can foster responsible model usage.
- Aligning Business and IT: maintaining enough flexibility in how policies are implemented throughout the organisation can help departments support their specific needs, and even aid in the value creation process itself.
- Rethinking the Separation between Management and Governance: Maintaining strong feedback loops between data management and data governance teams can improve processes in both camps, possibly in the form of a management-governance coupling CoE.
- Performance Management: Having quantifiable metrics of data governance success can help in both proving the value of data governance itself, and in providing feedback for future improvement towards the data governance effort.

Lastly, these questions then culminated in the final research question:

"What alterations need to be made to classical data governance to suit the needs of data governance in the context of Advanced Analytics?"

The answer to this final research question can be seen in a combination of both the above-mentioned solutions, and the governance framework discussed under "Governance Framework"(See Section 5.14).

Finally, the interview with Mr Snijders allowed the theories and concepts formed in this thesis to be grounded in reality. Although not all concepts discussed in the thesis translated directly towards Mr. Snijders and Alliander, the interview did provide some assurances that the core of these concepts is likely to translate towards real-word usage.

Although challenges within data governance on Advanced Analytics are likely to expand far beyond those discussed in this thesis (and may not show up until data governance is fully implemented), the discussed solutions and challenges combined with the governance framework, should give a strong indication of what factors carry importance in the setup of a data governance effort in the context of Advanced Analytics.

8 Discussion

This thesis set out to present an overview of the challenges found when applying data governance towards Advanced Analytics, combined with possible solutions for these challenges. Data governance has proven to be an interesting subject, requiring a multi-faceted approach that includes both technological, business and cultural aspects.

Due to the nature of an exploratory qualitative study, this thesis present no quantitative results to interpret and discuss. However, the findings from the literature study and the interview suggest that the significance of proper data governance should not be understated. Considering the impact that proper data quality and access have on the success of any data processing project, ensuring these factors meet the requirements set by organisation becomes a vital undertaking.

Although Mr. Snijders opposed some of concepts presented by the literature study - such as the importance of the Black-box problem and Model Selection - his views on data governance generally supported the findings, adding some level of validity towards the thesis. One thing that stood out from the interview was the focus on culture that Mr Snijders and this thesis shared. His views on data engineers taking accountability nicely lines up with the cultural aspects of accountability discussed in "Accountability in Usage" and "Data Lakes and Data Mesh Architectures". The growing understanding of the value that non-IT departments are starting to see in data governance also shows that the cultural swing necessary for data governance to work is also being seen in his organisation.

8.1 Future Research

Due to scope limitations, this thesis mainly focuses on more corporate-oriented organisations. However, these are not the only organisations where the rise of Advanced Analytics is likely to affect their operations. Other types of organisations will also be forced to make branch-specifics adjustments to how they approach their data:

- Education systems will not only use more technological tools in their teachings, but data about learning is also likely to have a more pronounced role. The US Department of Education has already invested in these technologies [73] to see how students move through learning trajectories, or how to create pathways through study material for specific learning objectives. Governing this specific data also requires a careful approach, considering that it may contain sensitive data about the students and teachers.
- Law enforcement has recently started using big data to assist in creating suspect profiles around crimes, such as identifying human trafficking networks in the US [73]. This pattern analysis can help avoid crime and stop criminals, but also uses personal information about large groups of people. Especially when data is not properly maintained and the inherent biases are misunderstood, existing biases about how law enforcement is executed can be reinforced, possibly leading to higher levels of discrimination. Properly managing how this data is used and collected is essential to protecting a nation's security and its civil liberties.
- Research centres use big data and Advanced Analytics for pattern analysis on patient data to find better ways of treating patients and developing medicine, but are

simultaneously working almost exclusively with sensitive data. Mismanagement of this data can not only violate privacy regulations of patients, but also lead to false positives in development of medicine, potentially harming people. Proper governance can prevent many of the risks within this field, and protect the reputation and research results from these centres.

Another factor is that due to this field being quite new and the thesis being of an early exploratory nature, it does not cover the full breadth of factors and variabilities that affect an organisation. For future research, one can delve deeper on a specific subject, or go wider and look at extra factors that were not discussed here. Some examples are to elaborate more specifically on details of data structures like data meshes by looking at their concrete implementations, or by considering how an organisations affects - and is affected - by an external organisations' data governance approach. One way to approach the exploration of new subjects would be to take the ideas suggested in this thesis, then incorporate them into a case study. This would allow one to consider which factor had the biggest effect on the organisation, so one can elaborate on that specific factor in future research.

Lastly, one major limitation lies in the fact that through unfortunate circumstances, this thesis could not be done at an external organisation, and was instead done internally at the Radboud University. Despite the addition of the interview, being able to test the elements against the inner workings of an actual organisation and forming a governance framework based on an actual case analysis would add some ecological validity to future projects. Governance frameworks and Advanced Analytics models in particular are things most organisation are rather secretive about, so having access to these could move some concepts in this thesis from conceptual blueprints to concrete implementations. Additionally, adding multiple interviews with different experts in the field would add further reliability and validity to the concepts discussed in this thesis, as it would highlight opposing and overlapping views across departments and organisations.

Due to these reasons, one of the main limitations of this thesis is the lack of validity. Whereas a more ideal situation would have had case studies or more extensive evaluations fill this validity gap, concrete validation here is limited to a single interview. If future research can avoid these limitations, there a few options on the table to improve upon the validity of this thesis.

First off, if one has access to the aforementioned internal elements of the organisation such as their governance framework or Advanced Analytics models, one can more concretely see whether and where elements in the organisation can be adjusted to better fit the needs of their Advanced Analytics. By adjusting their internal elements based on the proposed solutions from this thesis, one can see the actual effects of these measures on their organisation. This adjustment provides additional ecological validity, as it tests the possible solutions in a real-life environment. A rough approach to this method would be to:

1. Spend time within an organisation, learning to understand their general methodology, vision, processes and data governance culture.
2. Select one specified element of their organisation or data governance approach to adjust based on the challenges and possible solutions discussed within this thesis.
3. Apply the necessary changes, then select appropriate metrics for performance management to see how the changes affect the organisation over time.

The outcome of these metrics could then be used as a validation; improvements across these metrics would validate the impact of the possible solutions.

If one has access to multiple interviewees from either the same or different positions and organisations, one can:

- Conduct interviews with multiple members who share similar positions in different organisations. This would allow for comparisons of discussed concepts over multiple organisations, which should provide an extra level of validity and reliability. By comparing the outcomes from the interviews and seeing where their answer differ and overlap, one gets a better picture of the most general problems organisations run into, and thus what solutions present the most likely improvements within their governance framework. This would also allow for proper coding of the interview results, as there are now multiple interview to compare across. As mentioned, this is something which would have provided little value for a single interview. The general approach here would be to:

  1. Conduct multiple interviews with similar positions across different organisation, then set up the collected information to be processed by a coding tool such as Atlas.ti [74].
  2. Start the coding phase, which is about breaking down the text into separate codes that represent descriptions of your content. These codes are then grouped, which leads to organized categories of codes [75]. As an end results, this provides the patterns in theories in your interview, helping with the development of data governance theories and concepts.

  As these codes are based on experts from within the field, this categorisation provides additional ecological validity, with a broader reach than a single interview can.

- Conduct interviews across different positions- either within the same organisation, or over different organisations. As no department is the same, receiving multiple different views on the same problems is likely to provide very different answers across the board. Someone who works in HR may perceive very different elements of data governance as important compared to someone who works in finance. This allows for more specificity in approaching data governance across different departments. One approach here could be to:

  1. Conduct multiple interviews across these different positions, then see how element of the governance framework lines up with the provided answers for that specific organisation. Based on these answer, the framework can be adjusted appropriately.
  2. As an end product, this would provide a different variation of the governance framework for each interviewed position; one governance framework for HR departments, one for finance etc. These are all likely to share the same basic structure, but with specific elements emphasised and de-emphasised, based on the elements' importance for that specific position.

References

[1] B. Cheatham, K. Javanmardian, and H. Samandari, "Confronting the risks of artificial intelligence," 04 2022.

[2] "General Data Protection Regulation," 5 2016.

[3] B. Hufford, "The 4 Types of Validity in Research Design (+3 More to Consider)," 08 2022.

[4] International, DAMA, *DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition*. Technics Publications, second ed., 2017.

[5] J. Ladley, *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program (The Morgan Kaufmann Series on Business Intelligence)*. Morgan Kaufmann, 1 ed., 2012.

[6] Prukalpa, "Data Governance Has a Serious Branding Problem - Towards Data Science," 01 2022.

[7] International Organization for Standardization, *ISO/IEC 27001:2013, Second Edition: Information technology - Security techniques - Information security management systems - Requirements*. Multiple. Distributed through American National Standards Institute (ANSI), 2 ed., 2013.

[8] "A brief history of big data everyone should read," 05 2022.

[9] A. Ramzi, "Studying population of egypt based on census data and geographic information system," 10 2012.

[10] "IBM Archives: Herman Hollerith," 11 1972.

[11] Anaconda, Inc., "Anaconda — State of Data Science 2020: Moving From Hype Towards Maturity," 2020.

[12] Gil Press, "Cleaning big data: Most time-consuming, least enjoyable data science task, survey says."

[13] S. H. P. Aarnoutse, "Data Usership Under Governance: A Proposal For Data Governance Roles From A Usership Perspective Within The Context Of A Dutch National Bank," *Radboud University Master Thesis Lab*, 2021.

[14] B. V. Gils and V. H. Publishing, *Data Management: a gentle introduction: Balancing Theory and Practice*. Van Haren Publishing, 1 ed., 2020.

[15] B. Kantor, "The RACI matrix: Your blueprint for project success," 12 2021.

[16] R. Bose, "Advanced analytics: opportunities and challenges," *Industrial Management Data Systems*, vol. 109, no. 2, pp. 155–172, 2009.

[17] J. Ram, C. Zhang, and A. Koronios, "The implications of Big Data analytics on Business Intelligence: A qualitative study in China," *Elsevier B.V*, 2016.

[18] I. C. Education, "Unsupervised Learning," 03 2022.

[19] J. Wiener and N. Bronson, "Facebook's Top Open Data Problems," 10 2014.

[20] Splunk, Enterprise Strategy Group, "The Data Age Is Here. Are You Ready?," tech. rep., 2020.

[21] R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: A conceptual framework, structured review, and research agenda," *International Journal of Information Management*, vol. 49, pp. 424–438, 2019.

[22] IBM Institute for Business Value and Saïd Business School at the University of Oxford, "Analytics: The real-world use of big data," tech. rep., 5 2013.

[23] "IBM Developer," 03 2018.

[24] "Health Insurance Portability and Accountability Act," 1996.

[25] "Sarbanes-Oxley Act," 2002.

[26] "California Privacy Rights Act," 2020.

[27] "California Consumer Privacy Act," 2018.

[28] "Colorado Privacy Act," 2021.

[29] H. Sun and O. Inc, "Enterprise Information Management: An Oracle White Paper On Enterprise Architecture," tech. rep., 5 2011.

[30] P. J. M. van Laarhoven and E. H. L. Aarts, "Simulated annealing," *Simulated Annealing: Theory and Applications*, pp. 7–15, 1987.

[31] Y. Zhang, S. Wang, and G. Ji, "A comprehensive survey on particle swarm optimization algorithm and its applications," *Mathematical problems in engineering*, vol. 2015, 2015.

[32] O. Loyola-Gonzalez, "Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View," *IEEE Access*, vol. 7, pp. 154096–154113, 2019.

[33] J. Gryz and M. Rojszczak, "Black box algorithms and the rights of individuals: no easy solution to the "explainability" problem," *Internet Policy Review*, vol. 10, no. 2, 2021.

[34] Hd, "A Practical Guide to Interpreting and Visualising Support Vector Machines," 06 2022.

[35] A. Navia-Vázquez and E. Parrado-Hernández, "Support vector machine interpretation," *Neurocomputing*, vol. 69, no. 13-15, pp. 1754–1759, 2006.

[36] V. Van Belle, B. Van Calster, S. Van Huffel, J. A. K. Suykens, and P. Lisboa, "Explaining Support Vector Machines: A Color Based Nomogram," *PLOS ONE*, vol. 11, no. 10, p. e0164568, 2016.

[37] S. Chen, C. Gao, IEEE, and P. Zhang, "Enhancing Interpretability of Black-box Soft-margin SVM by Integrating Data-based Priors," *Cornell University*, 2019.

[38] S.-i. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, 1993.

[39] R. Pappadà and F. Pauli, "Discrimination in machine learning algorithms," 01 2019.

[40] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," 10 2018.

[41] ProPublica, "The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review," 02 2020.

[42] European Union Directive, "Racial Equality Directive, Art 2(2)(b)," 2000/43/EC.

[43] E. Pintelas, I. E. Livieris, and P. Pintelas, "A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability," *Algorithms*, vol. 13, no. 1, p. 17, 2020.

[44] J. Glanz, "Data Centers Waste Vast Amounts of Energy, Belying Industry Image," 09 2012.

[45] P. P. Tallon, "Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost," *Computer*, vol. 46, no. 6, pp. 32–38, 2013.

[46] D. R. Merrill, "Four Principles for Reducing Total Cost of Ownership," *Hitachi Storage Solutions*, 2011.

[47] L. Madsen, *Disrupting Data Governance: A Call to Action*. Technics Publications, 2019.

[48] M. Zetlin and T. Olavsrud, "What is a chief data officer? A leader who creates business value from data," 03 2022.

[49] D. A. Team, "CIO vs. CDO: Roles in Digital Transformation," 07 2021.

[50] O. Benfeldt Nielsen, "A comprehensive review of data governance literature," 2017.

[51] R. Abraham, J. Schneider, and J. vom Brocke, "Data governance: A conceptual framework, structured review, and research agenda," *International Journal of Information Management*, vol. 49, pp. 424–438, 2019.

[52] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy Artificial Intelligence," *Government Information Quarterly*, vol. 37, no. 3, p. 101493, 2020.

[53] C. Sorensen, "What is an Analytics Center of Excellence?," 10 2021.

[54] M. Rosemann, *The Service Portfolio of a BPM Center of Excellence*, vol. 2, pp. 267–284. 08 2010.

[55] Deloitte, " Evolving The Data Analytics Operating Model," tech. rep., 2020.

[56] J. Baijens, T. Huygh, and R. Helms, "Establishing and theorising data analytics governance: a descriptive framework and a vsm-based view," *Journal of Business Analytics*, vol. 5, no. 1, pp. 101–122, 2022.

[57] A. A. Avery and K. Cheek, "Analytics Governance: Towards a Definition and Framework," *Emergent Research Forum, Illinois*, 2015.

[58] K. Hill, "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did," 04 2016.

[59] C. Duhigg, "New York Times; How Companies Learn Your Secrets," 02 2012.

[60] H. Yeong Kim and J. Suh Cho, "Data governance framework for big data implementation with NPS Case Analysis in Korea," *Journal of Business Retail Management Research*, vol. 12, no. 03, 2018.

[61] "From data mess to a data mesh," 07 2022.

[62] L. Gavish and B. Moses, "What Is A Data Mesh — And How Not To Mesh It Up," 06 2022.

[63] J. Serra, "Data Mesh: Centralized vs decentralized data architecture — James Serra's Blog," 07 2021.

[64] E. Verheul, "Lecture 2: Information Security Risk Assessment," 2021.

[65] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California law review*, pp. 671–732, 2016.

[66] Centre On Regulation In Europe, A. de Streel, A. Bibal, B. Frenay, and M. Lognoul, "Explaining The Black Box: When Law Controls Ai," tech. rep., 02 2020.

[67] N. Tollenaar and P. G. M. van der Heijden, "Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 176, no. 2, pp. 565–584, 2012.

[68] C. Rudin, "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition · Issue 1.2, Fall 2019," 11 2019.

[69] I. Grau Garcia, D. Sengupta, M. Garcia Lorenzo, and A. Nowe, "Grey-box model: An ensemble approach for addressing semi-supervised classification problems," in *Proceedings of the 25th Belgian-Dutch Conference on Machine Learning BENELEARN 2016*, pp. 1–3, 9 2016.

[70] D. Consultants, "Data Governance Act: main elements and business implications," 12 2021.

[71] "Data Act: Shaping Europe's Digital Future," 07 2022.

[72] E. Verheul, "Lecture 10: Information Security Audits Certification, slide 7," 2021.

[73] Executive Office of the President, "Big Data: Seizing Opportunities, Preserving Values," *White House*, 2014.

[74] K. Kusi-Mensah, S. Poleschuk, and K. Riopelle, "ATLAS.ti — The Qualitative Data Analysis Research Software," 08 2022.

[75] "Make the Best of Codes in ATLAS TI," 07 2022.