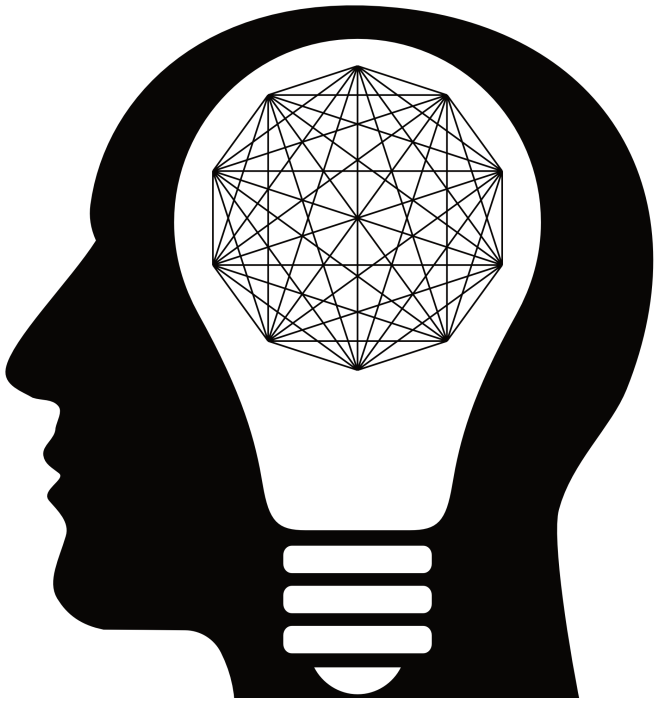
CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG



Requirements Engineering for Machine Learning

Dealing with the exploratory nature of ML development projects

Master's thesis in Computer Science and Engineering

Isac Boman
David Welander

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Master's thesis 2023

Requirements Engineering for Machine Learning

Dealing with the exploratory nature of ML development projects

Isac Boman
David Welander

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2023

Requirements Engineering for Machine Learning
Dealing with the exploratory nature of ML development projects
Isac Boman
David Welander

Supervisors: Jennifer Horkoff & Eric Knauss, Computer Science and Engineering
Industry Supervisors: Mats Honner & Mattias Jonhede, Volvo Group
Examiner: Birgit Penzenstadtler, Computer Science and Engineering

Master's Thesis 2023
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Head in profile with neural network and lightbulb, by Gordon Johnson via Pixabay

Typeset in LaTeX
Gothenburg, Sweden 2023

Requirements Engineering for Machine Learning
Dealing with the exploratory nature of ML development projects
Isac Boman
David Welander
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

While there are many well-established Requirements Engineering practices for traditional, deterministic, software systems, the emerging field of Machine Learning introduces new challenges with Requirements Engineering. Current theoretical Software Engineering research has identified many challenges with RE for ML, but there is currently a lack of empirical evidence. Challenges are thought to arise for example because of the uncertain nature of ML, and the dependence on data. Innovative ML development is also highly creative, potentially introducing a trade-off between requirements and creative freedom. Through a case study, based on interviews, observations, documentation, and a combined focus group and questionnaire, this thesis provides insight into what challenges and success factors related to RE for ML that are present in an empirical setting, and compares them to literature in the research field. The thesis further recommends that practitioners in the field use a variation of Goal-Oriented Requirements Engineering, ML-GORE, together with practices aimed at understanding the domain and user, such as use case diagrams and scenario-based requirements elicitation. It is also recommended that the stakeholders are involved in the entire requirements and development process. However, it is suggested that these practices, and their impact on RE for ML, are validated in further research. Finally, the findings confirm that new challenges arise when applying RE to ML development. These challenges are to a great extent in line with previous theoretical research, with two of the major ones being data dependence and outcome uncertainty.

Keywords: requirements engineering, requirements specification, machine learning, advanced analytics, creativity, software engineering, thesis, case study, RE for ML.

Acknowledgements

We would like to thank our academic supervisors Jennifer Horkoff and Eric Knauss for their everlasting support and positive spirit throughout the thesis. Further, we would like to thank our industry supervisors Mats and Mattias, as well as the entire Advanced Analytics department. It has been a pleasure to join the team and get to know everyone.

Isac Boman & David Welander, Gothenburg, May 2023

Contents

List of Figures

List of Tables

Abbreviations

Abbreviation	Description
AdA	Advanced Analytics
AI	Artificial Intelligence
DIM	Dealer Inventory Management
DIP	Demand and Inventory Planning
DoD	Definition of Done
GORE	Goal Oriented Requirements Engineering
GTO	Group Trucks Operations
ML	Machine Learning
NFR	Non-functional Requirement
RE	Requirements Engineering
SAFe	Scaled Agile Framework
SCAE	Supply Chain Analytics Expert
SCO	Supply Chain Optimization
SE	Software Engineering
SML	Service Market Logistics

1

Introduction

With the rise of Artificial Intelligence (AI) and Machine Learning (ML), researchers have recently begun to examine the effects of including ML as part of software development. ML is often part of larger complex systems, and most parts of such systems can follow standard development processes, for example, Scaled Agile Framework (SAFe), Scrum, or the V-model [1] [2]. These processes are well-defined and studied, but the inclusion of AI and ML presents new challenges to these development processes [3]. When it comes to ML-specific development, these methods typically fail to account for ML's explorative nature and non-determinism [4].

More specifically, Requirements Engineering (RE) for ML has been identified as a particularly problematic and challenging task in ML development [5]. This has been an area of focus in the RE community, with mainly theoretical analyses of challenges and potential solutions. However, the field is young and there is substantial work left to do, especially in terms of evaluation and practical approaches [6]. RE is a crucial factor when it comes to increasing the chance of project success in software development [7]. Further, the impact grows for larger projects or companies, when there is a customer requesting certain functionality or performance. Some requirements are said to be tacit, or intuitive, and are hard to formulate into requirements, but the other, explicit, requirements should always be documented to a certain extent.

In a case study by Salerno et al. [8], it becomes obvious that the traditional innovation process, linear from idea to launch, is not the only model of innovation. In fact, just under half of the studied companies use other means of handling innovation projects. They argue that the linear model is not suitable for technologically complex or uncertain innovation processes and that in most cases there has to be some form of tailoring of the process to the specific company or context [8]. Further, there are also challenges in what Pavitt refers to as *Heterogeneity in Innovation Processes* [9]. He argues that the search for innovation in a company is heavily influenced by its previous success stories, making disruptive innovation processes even harder. Pavitt also concludes that "there can be no simple 'best practice' innovation model for firms or managers to follow" [9, p. 96]. In line with the aforementioned conclusion that there are specific demands for innovation management for each innovation context, the rising field of ML development gives cause to analyze its processes for innovation. Not only is ML technologically complex, but its novelty also makes the innovation processes, and the potential outcomes, uncertain and highly explorative.

1.1 The Case

The case study will take place in the Advanced Analytics (AdA) team at Volvo Group. Organizationally it belongs to Group Trucks Operations (GTO), and more specifically Service Market Logistics (SML) - Supply Chain Optimization (SCO). The overall responsibility of the team is to use advanced analytics, for example, ML to optimize the pattern recognition and forecasting of the need for spare parts and services, in order to increase value to the end customer and reduce cost in the supply chain. AdA does not only work on projects within SCO but carries out projects throughout the entirety of Volvo Group where there are relations to spare parts and service. The output is software products on multiple levels, from single Python scripts to models that are fully integrated into the internal systems. The products are used for supply chain optimization purposes in the automotive spare parts sector, but the methods, frameworks, and packages used in the models are typically general for ML or advanced analytics. Examples include common Python packages like Pandas and Scikit Learn.

1.1.1 Stakeholders and Customers

As mentioned above, the AdA team serves a multitude of different customers internally at Volvo. Most of their collaborations have been with operational units in SML, but there have also been projects with other parts of Volvo Group, organizationally further from AdA. The degree of collaboration in an exploration can vary greatly, from stakeholders actively participating in development to placing a request for delivery. In general, the projects have been focused on forecast accuracy, data handling and visualization, and ML applications that automate tasks with the purpose of saving time. Two examples of partner organizations are described below:

Demand and Inventory Planning
Demand and Inventory Planning (DIP) are primarily responsible for forecasting the demand for spare parts at the central warehouse, but also warehouses closer to the end customer. Included in this is also analyzing and setting parameters on for example safety stock and economic order quantity. They have collaborated with the AdA team on different projects to enhance their use of data and advanced analytics to increase forecast accuracy, improve the handling of seasonal variations, and optimize previously mentioned parameters.

Dealer Inventory Management
Dealer Inventory Management (DIM) works closely with the dealers. DIM ensures that spare part stock levels at dealerships are in line with the service level agreements and that forecasts and restock plans are accurate and up to date. As an example, the end customer has different agreements with Volvo that ensure different availability of parts. This has to be taken into account when stocking dealership sites, while minimizing inventory and logistic costs. In collaboration with AdA, DIM focus on visualization of data and tries to build the grounds for data-driven decision-making.

1.1.2 Team composition

The AdA team consists of technical experts, domain experts, and managerial positions. However, people outside the team can take part in their development process, both as active stakeholders and as any of the roles mentioned above. There are also people dedicated to understanding and guiding the organization in the transformation towards analytics and data-driven processes. Figure 1.1 depicts the team and role structure. The team acts in multiple organizational contexts, with a local core operations team and a more outward-facing, global team.

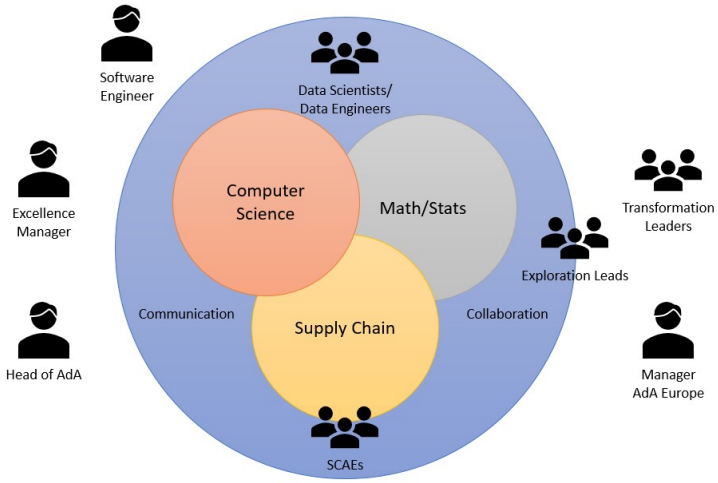


Figure 1.1: The Case team structure and competence areas as presented by the organization. The blue circle marks the distinction between the local and global team.

The Data Scientists and Data Engineers work, as can be expected, with the implementation of the analytical solutions. They process and prepare the data, build the models, and adjust them before handing the solution over to the customers, who are different operational teams within Volvo Group. However, the role also involves some software engineering, deployment, and maintenance of the models, which could be considered out-of-scope based on the role's title.

There are several domain experts, or Supply Chain Analytics Experts (SCAEs), who are responsible for translating between the customer and the implementation. SCAEs are often well-versed in development practices and often participate in the implementation as well.

The managerial side contains two levels, one being more project-oriented and one more staff and team-oriented. The Exploration Leads are in charge of coordinating, supporting, and managing individual explorations. They are leading stand-ups, clearing blockers, and representing the exploration in decisions. Further, there is one global and one local department head, who are the formal team leaders. Their

responsibilities lie further away from the operations and are more inclined toward the rest of the organization. They are responsible for the general progress of the team, and its strategies for the future.

Lastly, Transformation Leaders act as cultural representatives toward the larger organization. They are responsible for preparing and training the organization towards an analytical mindset and increasing user maturity and readiness for future projects. Similar to the department heads, they are less involved in the explorations, but they can also act to translate and identify barriers that hinder the implementation or transformation of the customers' operations or businesses.

1.1.3 Explore Mode Development

In line with Gartner's view on explorative processes [10], AdA has adopted the second part of bimodal development where traditional development and operations are complemented with a second mode, exploration mode. Gartner [10] refers to this as Mode 1, traditional development, and Mode 2, explorative development. Explore mode development is used for ML and other AdA ideas where there is no standard solution and many unknown factors. It is meant to foster creativity, exploration, and experimentation, and to shorten innovation cycles. The exploration mode process has adopted elements of design thinking [11], and hence, all explorations start with Empathy and Ideate and then enter a cycle of exploration. Empathy is a phase where more is learned about the problem and stakeholders, and Ideate is a brainstorming phase where ideas and possible solutions are investigated. The team's process, internally called "the trumpet", is illustrated in Figure 1.2.

Ideations and Empathic Designs are funneled into the trumpet from a wide variety of stakeholders. They can appear inside their own team when opportunities are found within technology or data, or a domain expert finds a new application area. Ideas can also emerge from external stakeholders. During these early phases, the team increases the understanding of the customers' problem and outlines the idea. They also perform a so-called "Quick evaluation" which serves as a short documentation of the preconditions of the idea on several factors, with each factor getting a ranking between zero and three. The factors are shown in Table 1.1. These ideas are stacked in an exploration idea backlog, from which the team selects projects to enter the exploration cycle. These selected projects then become so-called "explorations". The cycles can be compared to the regular agile notion of a sprint, but the number of cycle iterations per exploration varies greatly. At the end of each cycle, there is a decision meeting with three possible outcomes: reiterate the exploration in the next cycle, kill the project and formulate learning outcomes, or move the exploration into the pilot phase.

When an exploration becomes a pilot, it enters the real-world setting for the first time. The ML models or applications are user tested, and their value statements are defined with more precision. If the prototype creates the expected value it is scaled in order to finally reach the intended scope of implementation and enter a

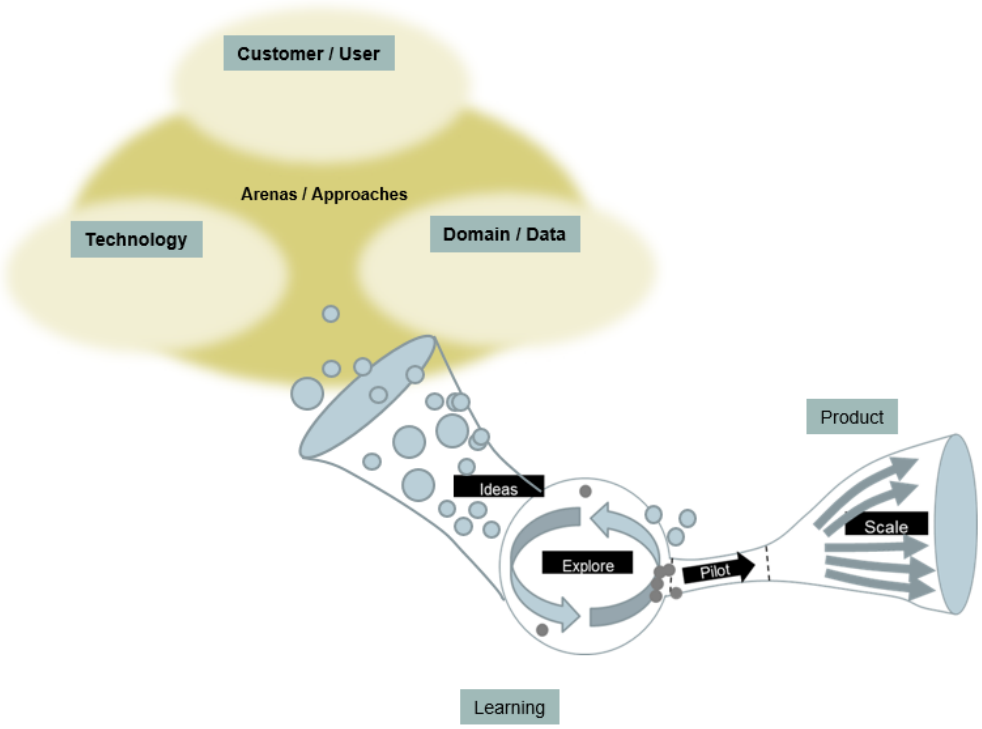


Figure 1.2: "The Trumpet" Explore Mode process, as applied in the case organization. *Volvo Group, 2022, reprinted with permission.*

Table 1.1: Quick evaluation factors. The scale of evaluation ranks between zero and three. Created by and used internally for AdA.

Factor	**Description**
Level of Impact	How much the company gains from doing a task.
Usability (data)	An estimation of with which effectiveness/efficiency the data can be worked with.
Accessibility (data)	An estimation of the effort on how to reach the data.
Level of existence (data)	An estimation of to what degree needed data exists at all.
Technical feasibility	An estimation of the technical risk.
Organizational and people change	An estimation of willingness and ability to assimilate an advanced analytics solution.
Sustainability impact	An estimation of the sustainability impact of a task, relative to today's solution.
Originality	A subjective estimation of the new combination of data, domain, technology, or concept.

state of operation and monitoring. The aforementioned value estimations follow a step-by-step refinement process, as pictured in Figure 1.3.

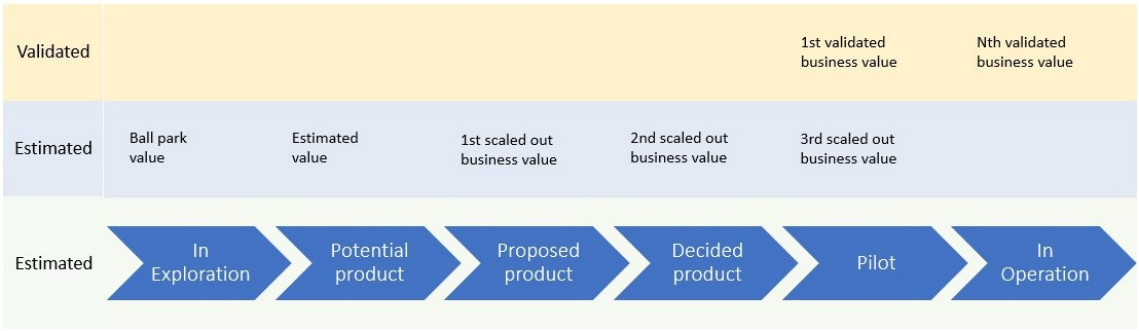


Figure 1.3: Value estimation refinement as carried out through explore mode development.

1.2 Purpose of the study

Exploration mode largely relies on a bottom-up approach where ideas rarely emerge from management, but from all levels of hierarchy. Due to this spontaneous generation of ideas, there is often little knowledge of the predicted outcome or implementation when starting out. As of today, creativity is central to the process and as such, the amount of management decisions is limited in the early stages of exploration. The data and different opportunities are explored, business knowledge is considered throughout the process, and different ML models arise in a very creative and untethered way. However, even these early exploratory processes need some structure, more specifically in terms of RE, since it is part of a larger organization in need of a common direction and coherence between departments. The problem at hand is to identify RE challenges, evaluate current practices, and analyze whether there are existing RE models or techniques that can be reused and applied to the relatively new and unexplored field of RE for ML.

The purpose of the study is to provide empirical case evidence to contribute to the existing body of knowledge on RE, agile RE, and RE for ML. The purpose is further to identify suitable RE practices in other fields of SE, and suggest them for further validation in the field of ML development. By performing a case study, theoretical frameworks can be evaluated and compared to practical evidence, which can be beneficial both for practitioners or organizations considering RE and researchers by expanding the empirical knowledge in the field.

1.3 Research questions

Considering the problem and purpose above, the research questions (RQs) will focus on RE for ML, mainly in the Explore Mode context. The four research questions are formulated below.

RQ1 Which are the major observable challenges present in RE for ML processes?
RQ2 How do the practically observed challenges compare to the theoretical challenges?
RQ3 What are the characteristics of successful and unsuccessful ML projects and processes?
RQ4 What commonly used techniques and practices for RE could be suitable for ML projects?

The first three research questions identify the challenges and characteristics in RE for ML, but to find a suggested practical solution to these, the fourth RQ is introduced to investigate models, techniques, or practices that could possibly solve the identified challenges.

2
Background

In this chapter, a review of the current state of research relevant to the study is presented. An introduction to the evolution of the software engineering discipline is given, together with relevant findings that concern SE and RE aspects of ML. Finally, a summary of the challenges found within RE for ML is presented, divided into six high-level groups.

2.1 Traditional SE and RE

The traditional SE approach was, to some extent, to utilize engineering principles on the previously more unstructured software developing process [12]. This meant a great deal of effort was put into planning and specifying requirements before the actual development started. This has in many cases been visualized with the so-called waterfall model. As the name suggests, the model is sequential and gated, with the significant effort required to move back to a previous step [13]. As a predictive model, the scope and initial specification of the product are key, since there is typically no user interaction during the development process. The finalized product is then handed over to the customer at the last stage, hopefully, according to the initial specifications. However, as Flewelling [13] describes, the longer the time frame, the greater the risk of changing business states and requirements.

The traditional RE practice is, in short, to produce a comprehensive requirements artifact before any design or development, and to use this during the process. While requirement management is mentioned, for example, handling changing requirements, the traditional model is largely defined from the assumption of a perfectly sequential development process [14]. However, this is seldom used in practice [14]. The traditional requirements artifacts can be divided into five parts according to Lauesen: an introductory description, system limits, data requirements, functional requirements, and quality requirements [14, p. 32]. According to the sequence, these are defined before designing and implementing any parts of the system. Once again, this theoretical separation between the steps is hard to find in practice.

There are some innate challenges with the traditional approach, most of which are connected to the knowledge and understanding of the stakeholders. The stakeholders might not have figured out what they want, or lack the understanding to express it adequately. The perceived requirements might also not be what the stakeholder wants in the end, after reviewing prototypes. Further, if the product is innovative,

there might not even be enough knowledge of the domain to express any functional requirements at all except perhaps the general business goals or values [14]. As ML can be considered such an innovative field of development, the traditional approach might be limited, and a more Agile RE approach should potentially be applied.

2.2 Agile SE and RE

Given the challenges posed in section 2.1 on traditional RE, there is a need to understand the more flexible RE processes within agile RE. Furthermore, this literature is highly relevant to this thesis since the case organization aims to apply an agile SE process and hence should approach RE from an agile perspective. In a systematic literature review by Inayat et al. [15] 17 key elements of agile RE are listed, of which the five most mentioned were:

- Requirements Prioritization, which differs from traditional RE in the fact that it is carried out iteratively over the entire development process, instead of in the beginning.
- Testing before coding, which is a typical agile practice, which connects the development to the functional requirements and creates direct feedback loops.
- Face-to-face communication, which reduces the need for written requirements and further increases the flexibility of the requirements process.
- Customer involvement, connecting to the aforementioned face-to-face communication was stated as "the primary reasons for project success and limited failure" [15, p. 921]. Physical proximity to the customer was also mentioned.
- Iterative requirements, that are evolved close to the development, throughout the process, and prioritized continuously as mentioned above. The iterative mindset is present throughout the agile practices.

Though Inyat et al. argue that agile RE is on track to meet some of the major challenges of traditional RE, they also conclude that there is a need for further research in the area. More specifically the paper suggests "further empirical evaluation of practices in industrial cases" [15, p. 926], which is connected to the purpose of this thesis, particularly for ML.

Similar statements of agile RE practices can be found in an article by Cao and Ramesh [16], where a study of 16 organizations resulted in seven key practices: Face-to-face communication, Iterative RE, Extreme prioritization, Constant planning, Prototyping, Test-driven development, and Reviews & Tests [16]. The article gives no recommendations for further research but presents challenges with each of the seven practices, and organizations are recommended to analyze the cost-benefit trade-off before applying agile RE.

Kasauli et al. [17] present 24 additional RE challenges for large-scale agile development, grouped into six areas. The area "Support Change and Evolution", is considered to be especially important as background to this thesis. In that area, the they addresses the challenge of integrating experimental requirements, development,

and prototyping in the larger agile process of a company. The proposed solution to the problem involves handling requirements as artifacts linked to the product or experiment, but states that the overall handling of these separate requirements remains to be solved [17]. Kasauli et al. also argues that more focus is needed on the processes surrounding updates of requirements, and how that knowledge is transferred through the organization. It concludes that "more research on these aspects is urgently needed to provide better guidance, approaches, and tools to manage evolving requirements." [17, p. 22]

2.3 SE for ML

Incorporating ML components into large and complex software systems presents new challenges that cannot be addressed using traditional software development processes. In recent years, there has been a growing interest in this area, as indicated by the significant increase in publications since 2015, as shown by Nascimento et al. [3]. In an article by Giray [18], the findings are organized into the main areas of traditional SE, with the "Testing and Quality" and "Software Development & Tools" areas attracting the most attention. However, challenges have been identified in all areas, and the list of these challenges is extensive. Moreover, the study suggests that there is a lack of tools and processes to tackle these challenges and that industrial viewpoints should be considered in future studies. Additionally, the novelty of this field, which is evident from the fact that most papers studied by Giray [18] and Nascimiento et al. [3] was published recently, adds to the difficulty as there are few techniques and practices available to deal with these challenges.

The challenges that arise are to a great extent due to the specific characteristics of ML and its processes regarding development. One of these characteristics is the uncertainties involved in ML development. Giray [18] found that compared to traditional SE, it is more difficult to predict to what extent ML can solve a given problem. The uncertainty also affects the ability to plan the development and estimate the effort and resources needed [19].

To continue on how ML differs from regular SE, Amershi et al. state: "First, Machine Learning is all about data" [20, p. 1]. In difference to regular software engineering, the quality and characteristics of the data are foundational for the quality of the model, making ML processes inherently data dependent. Articles by Amershi et al. [20] and Wan et al. [21] argue that the success of an ML system relies, to a great extent, on having relevant and qualitative data available. This poses challenges as there is a lack of processes and tools for managing data throughout the development. As an example, Arpteg et al. [19] mention that specific data testing could be a measure to ensure data quality, but there are few tools available for this.

An often prioritized quality attribute of software is modularity, a measure of the coupling between different components, which contributes to the overall maintainability [22]. Amershi et al. [20] and Arpteg et al. [19] argue that this is more difficult to achieve in ML systems. According to Belani et al. [23], the low modularity results

in a phenomenon called CACE (Change Anything Change Everything). In addition to the internal coupling within the ML system, Arpteg et al. [19] also presents the potential challenge with external coupling, as ML systems often are dependent on several other systems or are incorporated into a large ecosystem.

2.4 RE for ML

Applications that have previously depended on the skill of the developer, UX designer, or other roles, are now more dependent on for example the quality and robustness of training data and input data [24]. This poses new challenges for RE from a number of perspectives. It is suggested that Data Requirements should be an entirely new category of requirements specifically for ML systems, including for example security, robustness, reliability, update intervals, quantity, and diversity [25]. Wan et al. [21] also found that software practitioners noted that the entire RE process is more data-driven and that requirements change depending on the systems' underlying data. The study also states that understanding the data, its features, and its distributions is key to ensuring proper quality assurance of the data and, as a result, the model. A recent literature study by Nascimento et al. [3] also finds that data management is a highly present challenge in studies on ML development, which further support the need for improved RE practices to ensure data quality. Lastly, there are challenges relating to requirements on, and practices for, data version control, which if handled incorrectly might impact the reproducibility and transparency of the development process [19].

To understand the data, one also has to understand the context or the domain from where the data is gathered. The meaning of certain features, a potential skew in the distribution, and over- or under-representation can be better comprehended if there is sufficient knowledge of the domain [21]. Belani et al. [23] also found this to be a challenge, especially in the elicitation phase of RE. It is not only important to understand the data, but also because the implementation of the model, and the requirements it should fulfill, is highly dependent on the domain. This is in line with Heyn et al. [24], who argue that the domain or context, in which the model is to be developed and deployed, should be specified. The main argument for this is that the performance of the model can change drastically if it were to be deployed in another context than it was developed for.

Another effect of the data dependency of ML, brought up in several papers, is that there is a great possibility of degradation in performance over time. Vogelsang and Borg [25] highlight the fact that ML regularly needs to be retrained, and Artpeg et al. [19] add that this maintenance of the model requires plenty of resources. Wan et al. [21] argue that the degradation could, to some extent, be predicted and should therefore be accounted for in the requirements processes.

The feasibility of an ML model is difficult to estimate early on in a process, as there exists a large degree of uncertainty in the development. One goal of specifying requirements for software is to align the parties involved on what the problem is and

how should it be solved. Both Ishikawa and Matsuno [4] and Arpteg et al. [19] argue that the uncertainty makes this process more difficult. This is a characteristic of ML that heavily affects RE, as it primarily takes place before the development starts. It adds a level of uncertainty to the requirements and is something that has to be dealt with [21]. In another study by Ishikawa and Yoshioka [5], "Lack of oracle" was the most frequent mentioned cause of difficulties in ML engineering, in a survey of 278 practitioners. It is perceived to be difficult to define any criteria or requirements regarding the output of ML systems. The unpredictability also introduces a challenge in the collaboration with stakeholders. The unpredictability, together with the general complexity of ML, can make it more difficult to communicate with stakeholders and manage their expectations [21]. There is a risk that the stakeholders have an inflated perception of what ML is capable of, and therefore believe that it can solve any problem [18]. This introduces a potential gap between the developers and the stakeholders that need to be handled [5].

The field of RE has primarily been established for traditional software products and includes practices, types of requirements, prioritization of requirements, and the trade-off between them. As ML differs in several ways, adaptation and rework are needed in RE for ML. One of the findings by Villamizar et al. [6] was that there is a lack of validated techniques applicable in RE for ML. Horkoff [26] focuses specifically on the Non-Functional Requirements (NFRs) and argues that ML NFRs differ from regular NFRs, and that there is a lack of knowledge regarding quality characteristics and potential trade-offs between said characteristics. NFRs are further investigated by Habibullah et al. [27], who agree with previous findings that NFRs differ for ML development. It is also found that there are challenges in measuring some ML NFRs, for example, explainability. Wan et al. [21] highlight that functional requirements, which are often the main type of requirement in traditional RE, do not have the same role in RE for ML. Instead, the use of quantitative measures is more dominant. There is also a greater focus on the ethical aspect of ML. As the models can be used in decision making, and in many cases rely on personal data, there is an increasing demand that, for example, discrimination and privacy are taken into consideration [25] [19]. NFRs like integrity and security are also found to be of high importance for ML systems [27].

ML development is usually done by data scientists or by more specialized roles, for example, ML engineers. These roles do not usually have the same background and experience as software engineers, or other roles that usually take part in software development, and there can be cultural differences. Both Amershi et al. [20] and Artpeg et al. [19] argue that this could pose a potential challenge with collaboration. In the software engineering discipline, the structure and quality of the code are prioritized, but this is not always the case for ML. As the two fields differ in several ways it also requires the involved parties to have a broader knowledge base. Wan et al. [21] argue that requirement engineers need to adapt to this and build a good technical understanding of ML, to be able to correctly elicit and formulate requirements. Nascimento et al. [3] instead focus on managing roles, and the importance of that they have good knowledge of both fields in order to facilitate communication

and collaboration, thus minimizing the risk of challenges due to cultural differences.

Further issues arise when regarding comprehension, which in this case could include both comprehension of the problem, the techniques, and how to comprehend and interpret the outcome of the ML models. ML models are inherently difficult to comprehend, as described by Arpteg et al. [19]. In order to improve the accuracy of models, the trade-off has been to enclose their inner workings, the training, in end-to-end black boxes. This is further discussed by Belani et al.[23] when trying to establish a taxonomy of ML challenges in RE. They argue that the black box characteristics hinder traceability and comprehension of how changing requirements alter the outcome. Relating to this, Heyn et al. [24] presents human factors and trust as a challenge that should be investigated further. The understanding of ML decisions, for example, in driver support systems, could impact the way results and decisions are being acted upon, or overridden, by humans. Ishikawa and Yoshioka [5] further conclude that "non-technical individuals are likely to be confused by the various streams of information on general Artificial Intelligence, human-beating game players, and so on." [5, p. 8], and that proper training and education have to be ensured in order to increase customer comprehension of ML systems.

How to measure the performance of a stand-alone ML model is relatively well understood. There are several metrics, for example, confusion matrices and ROC-curves. The difficulty arises when evaluating what the level of performance means in a given context. Vogelsang and Borg [25] argue that this should be done together with the stakeholders, but notes that it could be a challenge to translate the demands from stakeholders into performance scores. Ishikawa and Yoshioka [5] also found that the process of evaluating ML models together with the customer is different compared to traditional software. It is a more continuous and cyclic process as it is difficult to set expectations on the feasibility of the outcome in the early stages. In terms of metrics, Horkoff [26] adds that there is a general lack of understanding of how to measure quality attributes, or NFRs, for ML.

2.5 Challenge mapping

In order to compare challenges in theory with practice, the challenges were given a common name, and then grouped into six high-level categories. Table 2.1 shows the challenges and groups, as well as which sources mention the challenge. *Multi-disciplinarity* groups the challenges that stem from the increased need for different competencies in an ML project, which is expressed as challenges in collaboration between stakeholders, as well as in ensuring the cross-functional knowledge in the team. Since the field of ML, and RE for ML, is relatively new, the *Field Novelty* itself brings a number of challenges. There are no or few established practices, ethics and legal issues arise in this new field, and the conventional set of requirements is challenged. Further, the *Uncertainty* of ML SE transfers into RE for ML. The low predictability of the outcome makes RE difficult, and customers do not know what to expect from a system that might or might not be accurate or functional.

The literature states that *Coupling* of ML projects are harder to manage and therefore requirements on internal and external coupling might be a challenge worth examining. *Data Dependence* groups the challenges that are related to the need for high-quality data. This is a challenge in itself but it also relates to the understanding of the context from where the data is derived, challenges regarding the maintenance of the model, and the need to specify the context in which the model is to be deployed.

Lastly, challenges concerning *Comprehension* are grouped. If users cannot properly comprehend an ML solution, the acceptance of its implementation and outcome will be challenging. The choice of acceptance criteria and metrics, to evaluate the implementation success, is made more difficult by the relativity of what is a good implementation. For example, even models with low prediction scores might be better than nothing and the evaluation must be made on a case-by-case basis. The challenge of properly defining the problem is also linked to comprehension since the users or customers might lack the terminology or knowledge to properly express the problem and how it relates to ML.

Table 2.1: Challenges in, or related to, RE for ML.

Group	Challenge	Source
Multi-disciplinarity	Cross-functional Knowledge	[20], [23], [3], [21]
	Collaboration Difficulties	[19], [20], [25]
Field Novelty	Lack of Techniques & Practices	[6]
	Ethics & Legal Impact	[19], [3], [25], [27]
	Req. Types and Relations	[18], [26], [6], [21], [27]
Uncertainty	Outcome Unpredictability	[19], [18], [4], [5], [21]
	Customer Expectations	[18], [5], [6], [21]
Coupling	Increased External Coupling	[19], [26]
	Increased Internal Coupling	[20], [19], [23], [21]
Data Dependence	Evolution & Retraining	[19], [26], [25], [21]
	Data Quality Issues	[20], [19], [24], [3], [25], [21]
	Domain Knowledge Dependence	[23], [21]
	Domain Specific Viability	[24]
Comprehension	Achieving User Acceptance	[24]
	Evaluation, Acceptance & Metrics	[24], [26], [4], [25]
	Missing Problem Definitions	[23]
	Model Interpretation Difficulties	[3], [25]

3
Methodology

The questions considered in this thesis are highly contextual, and it is hard to distinguish a single phenomenon under study to separate from said context. Höst et al. [28] argue that for such questions, a case study could be suitable. Further, there is a need to "design for flexibility" [28, p. 23] in order to react to changes in the context during the case study. In fact, the researcher should assume that there will be change. In this case, a round of interviews was held, but the need for further confirmation arose, leading to both a focus group and a questionnaire.

Höst et al. [28] also argue that triangulation, for example, the use of multiple perspectives or sources, is important to conduct a valid case study. Therefore, this thesis aimed to examine the context and the ML project structure using interviews, observation through participation and reading current documentation of the case's processes, and validate this data through a questionnaire and a focus group (Methodological Triangulation). Multiple interviewees ensured Data Triangulation and the thesis also incorporated Observer Triangulation and Theory Triangulation [28], with two observers and different theoretical frameworks for RE in both agile and traditional software projects. The explore mode development process is the main unit of analysis, and multiple iterations of the process were considered within the same case context. This touches on the notion of separate units of analysis as defined by Höst et al. [28, p. 27]. Hence, an analysis can be carried out on the context, together with separate analyses of the units, in order to determine the traits and success factors of the units in comparison to each other. The process cycles are documented, which enabled a study of past cycles in addition to observation of ongoing ones.

3.1 Interviews

Interviews were carried out with the purpose of understanding the current state of processes and collaborations with stakeholders, as well as the interviewees' perceptions of current challenges in the explorative ML development process or specific projects. The format was that of a semi-structured interview, in order to let the interviewees identify and elaborate on characteristics of their involvement in explorations, as well as characteristics of successful and unsuccessful explorations. The complete interview guide is enclosed in Appendix A. After an interview, the audio recordings were transcribed and the transcriptions were sent out to be reviewed by the interviewee. No interviewees returned any changes to their transcripts.

The interview guide contained questions divided into five main topics: Interviewee Information, Case Context, Ongoing Projects, Past Projects, and Challenges in RE. The first two interviews also included an Interview Reflection section regarding the clarity of the questions, as well as whether the interviewee thought that the questions were leading. Initially, these interviews were to be considered pilots, and the data would be disregarded if there were changes to the template. However, there were no significant comments or critiques of the interview guide, and it was kept unchanged. Hence, data from the pilot interviews remaind valid.

3.1.1 Interviewee Sampling

The decisions when sampling interviewees was mainly affected by two factors, the size of the team and the interviewee's relations or involvement with the team. Interviews were held with all members of the local AdA team, which meant that no interviewee prioritization had to be made when considering the team members. The decision to interview all members was due to the relatively small team size.

There were also interviews with external stakeholders, which had to be selected and prioritized. In line with Palinkas et al. [29], purposeful sampling was used, based on the interviewees' knowledge and experience, to increase the chance of valuable information being elicited from the interviews. The selection was based on finding external stakeholders with recent involvement in an ML exploration, and with some geographic and organizational proximity. This was done together with our industry supervisors, who have great knowledge of people involved in explorations, to ensure that the interviewed stakeholders were relevant to this case. The interview template was kept unchanged for these interviews, since the questions were kept at a high level and the interviewees could answer all questions from their perspectives. The different roles interviewed are presented in Table 3.1, along with the number of interviews per role.

Table 3.1: Interviewee roles and counts. Roles are grouped by category to achieve anonymity.

Role	Relation	Count
Technical Expert	Team Internal	4
Domain Expert	Team Internal	2
Domain Expert	Team External	3
Management & Process	Team Internal	3

3.2 Observations

In order to strengthen the data collected from the interviews, observations were made throughout the case study. The authors were present in AdA's office for the duration of the study and participated in a number of forums, decision meetings, and presentations, where any observations were noted.

The observations followed a template in order to improve the analysis and comparison of the observations. An example of an observation is shown in Table 3.2. Note that some fields are omitted in the example. The complete observation template is enclosed in Appendix B.

Table 3.2: Observation template example.

ID	Context	Actors	Observation	Sentiment	Comment
O1	Introductory Presentation	Team lead	There's a decision gate between the Explore cycle and pilot. Criteria are not clearly stated.	Negative	Investigate criteria.

3.3 Documentation

The project documentation from past and ongoing explorations was collected in order to measure compliance with some current practices in the development team, specifically Quick Evaluation and ballpark value estimation. The documentation collected consisted of Kanban cards created in the team's Azure DevOps board. Each exploration gets its own card describing the project including information on, for example, Quick Evaluation.

Several data points from each card were extracted from Azure DevOps, listed in Table 3.3. All cards created between 2020-10-30 and 2023-04-11 were included, resulting in a total of 302 cards. After extraction, compliance with including a ballpark value was measured by comparing the number of non-null values with the total number of projects. Further steps were required when measuring Quick Evaluation compliance, where the value for all parameters was summarized and a sum of 0 was considered null (non-compliant).

3.4 Focus Group

As a second round of data collection, a combination of a questionnaire and a focus group was held. The questionnaire was presented and answered in the beginning of the session, followed by a focus group discussion, as described by Höst et al. [28]. The questionnaire served as a quantitative data source and as a basis for further

Table 3.3: Data fields extracted from project documentation and planning system.

Field	Description
ID	The exploration ID
State	The current state of the project
Work item type	Exploration filter parameter
Title	Exploration title
Accessibility	Part of quick evaluation
Level of existence	Part of quick evaluation
Level of impact	Part of quick evaluation
Organization and people change	Part of quick evaluation
Originality	Part of quick evaluation
Sustainability impact	Part of quick evaluation
Technical feasibility	Part of quick evaluation
Usability	Part of quick evaluation
Ballpark value	Estimated project value on completion

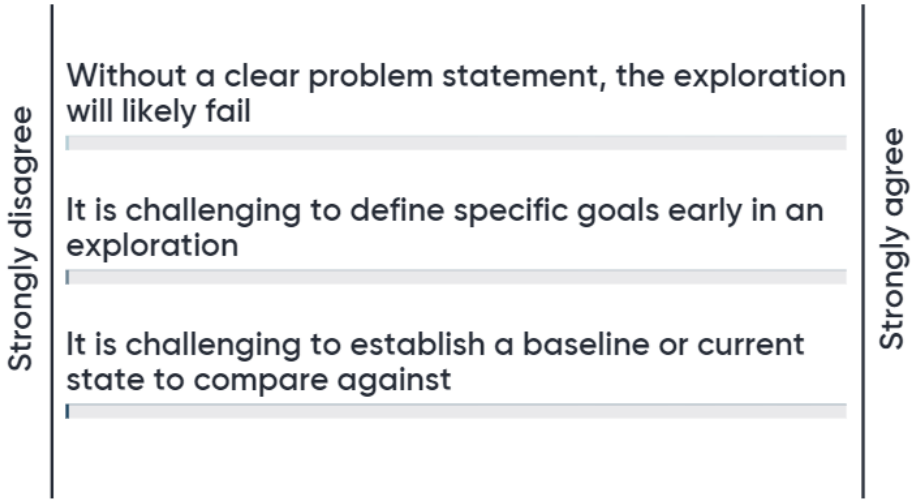


Figure 3.1: Example of questionnaire slide during the focus group session.

discussion in the group, facilitated by the authors. The questionnaire consisted of 25 statements divided into 9 topics and was derived from the primary results from the first round of data collection. An example from the questionnaire is presented in Figure 3.1, and the full list of statements can be found in Appendix E. The participants were asked to agree or disagree with the statements using a four point Likert scale, including the options "Strongly disagree", "Disagree", "Agree", and "Strongly Agree". The participants also had the option to not have an opinion on the statement. Any statements the participant found unclear were clarified during the session.

Table 3.4 depicts the participants in the focus group, anonymized in the same way as the interview participants list in Table 3.1. All interviewees were invited, and the group was expanded by inviting further members of the global AdA team that has a relation to the exploration process in the study, for example, global management.

A total of 6 invited participants could not attend the focus group, and they were asked to respond to the questionnaire as soon as possible after the focus group. Two participants joined late, both in the Management & Process category. They did not answer the questionnaire but participated in the following discussion. Their answers to the questionnaire were not collected afterward as they had seen the preliminary result. In total, 11 participants answered the questionnaire, of which 2 participants answered remotely after the focus group.

The purpose of this was to confirm the initial result gathered from interviews and observations. Through the questionnaire, quantitative data on agreement with our findings were obtained. Further, a focus group can support or deny statements by a single interviewee, which otherwise would be hard to draw conclusions from. The purpose of the focus group was also to get a deeper understanding of the themes, potential causes for or solutions to mentioned challenges, or other aspects that needs to be accounted for.

Table 3.4: Focus group participant roles and counts. Roles are grouped by category to achieve anonymity.

Role	Relation	Count
Technical Expert	Internal	4
Domain Expert	Internal	2
Domain Expert	External	2
Management & Process	Internal	3

3.5 Analysis

In order to process the interview transcripts and observations in a structured way, the texts were coded, as proposed by Höst et al. [28]. The coding was structured in two cycles, as suggested by Saldaña [30], which is suitable for a qualitative study. For the first cycle, the main method applied was Provisional Coding. A part of the study's purpose is to compare findings to previous research, which is arguing for the use of this method, where an initial set of codes is derived before data is analyzed [30]. The initial set was based on the research questions and the theoretical background, mainly revolving around three high-level topics: current practices, challenges, and success factors. Each topic could then have one or several subcodes. For challenges, the codes were based on Table 2.1, which was derived from previous research. The codes for practices and success factors were based on the research questions and understanding of the organization, and was initially kept on a high level letting the potential subcodes emerge during the coding. The initial set served as a starting point but not a finalized code set, as codes were added, changed, and deleted during the process.

The code book consisted of mainly two types of codes: attribute and structural [30]. Attribute coding was used on the meta-data gathered, for example, the role or the

experience of the participant or the location of the observation. Structural coding was used on the rest of the codes. The code, a short abbreviation, was used as a descriptor for the content of the coded text snippets. The complete code book is included in Appendix C.

Since a text snippet can have multiple analytical contributions, it could have multiple codes. For example, a snippet describing a success factor could also, in some cases, be a challenge. Saldaña [30] labels this as Simultaneous Coding. He mentions though that it should be used with caution, as it could be a sign of vague codes and could make the analysis more difficult. To keep track of this, the number of codes attributed to the same snippet was counted and in summary, 49 out of 516 text snippets were given multiple codes. In order to separate duplicate snippets from each other, the codes were added as separate rows in a spreadsheet, rather than having two codes on the same snippet. This simplified filtering, grouping, and analyzing the data.

The first cycle consisted of several iterations over the gathered data. The first iteration aimed to code all gathered data on at least a high-level topic, add potential codes to the set, and gain an initial understanding of the data. The coding was done simultaneously and independently by the authors for each snippet to increase reliability. Each author suggested a code for a text snippet and upon disagreement, the section was discussed. However, no agreement scores or measures were documented. After the first iteration, the code book was analyzed and modified. Each code got a clarified description of what it meant, the grouping of codes was revised, and unused codes were removed. One example was creating a clear distinction between *comments on* current practices and *suggestions for improvements in* current practices. Improvements were treated as a stand-alone subcode to practices. The goal of the second iteration was to establish stringency in the coding and correct potential errors made.

3.5.1 Further analysis

For the second cycle, the goal was to identify larger themes in the codes, that then could serve as the basis of discussion. This is what Saldaña [30] refers to as Pattern Coding, a method appropriate for grouping together codes into construct and themes, and as a means to find connections and explanations in the data. To facilitate this, all text snippets belonging to each code were summarized to condense the content of it. These condensations then served as a workshop base where the authors could point out common themes among the codes. After constructing the themes, these were verified against the low-level text snippets to make sure that the themes were in line with the actual answers and not the condensations. Further verification was also achieved in the focus group, see section 3.4.

One part of the analysis was a frequency diagram. This is a visualization of the number of interviewees that mention each category. As Saldaña [30] describes, this could be especially useful for structural codes to give an indication of the relevance

or occurrence of each category. This was used on both challenges and success factors, to highlight which are the most important ones and, in the case of challenges, compare against literature. To further substantiate the conclusions to be drawn from this, it was compared to the data gathered from the questionnaire.

The focus group discussions were recorded and transcribed. The transcript was then analyzed against, and grouped into, the analytical themes derived from the interviews and observations. These focus group statements could then be compared to the other data sources, for further triangulation.

4
Results

In this chapter, the results of the data collection are presented. First, a description of the case organization's current practices is given, followed by a presentation of the themes derived from the different data collection processes. Interviews, observations, and documentation examination are distinguished from the focus group and questionnaire, in order to highlight the level of agreement between the two data collection rounds.

4.1 Current practices

Aside from the introductory description in Chapter 1, this section will describe the case organization's practices as found in the data collection. These might differ, as chapter 1 states processes as described on paper, while the results show processes as perceived in interviews, observations, and documentation.

4.1.1 Project and team structure

The exploration projects in the AdA team are mainly generated in two ways. One is when the idea or problem is clearly defined by, or in collaboration with, stakeholders, and the other is a more explorative way of testing, for example, a new technique or package. The former is rooted in a problem, but for the latter, the team usually tries to find a use case to explore. The use case can then be creatively defined together with a stakeholder, to find a test-bench domain for the new technology.

One of the focus areas within AdA is creating a culture of explorative thinking. People should dare to try new techniques and methods, and are encouraged to take new approaches to problems. It was mentioned that this is largely dependent on leadership. It requires freer, less hierarchical leadership, but that does not exist in all parts of the organization. It is also mentioned that to facilitate explorative thinking, there has to be a balance focused between value creation and learning. Learning should be a part of every new project, and in some projects that could even be the main goal. This could slow down development, but at the same time, it builds important competencies both within the team and in the larger organization.

The level of exploration is higher in the early stages of a project, and the requirements are fewer or less specific. It is then gradually increasing, but it is not a clearly structured process of how and when. It was mentioned that too much structure both

in terms of requirements and processes, especially in the early stages, might hinder creativity leading to important aspects or values being missed. This is also said about the goal that explorations should take 30 days.

Regarding different modes of development, interviewees find that the team is not very good at differentiating between Mode 1 (exploiting) and Mode 2 (exploring). Usually, all ideas are implemented through Mode 2, but it was mentioned that there could be cases where the explorations should be less explorative and treated more as regular projects (Mode 1). Interviewees said that most explorations are unique. There is a formal structure that is the same for all explorations, but it differs in practice. Mostly, the team consists of at least one data scientist, one SCAE, and one exploration lead. Sometimes a stakeholder is part of the team as well. It is mentioned that explorations often lack ownership and that projects lose structure since prioritization is mostly interest-based.

4.1.2 Goals

Some interviewees said that explorations start with a discussion around the value statement and what goals to reach. However, other answers that vague goals appear during the exploration cycle. There are suggestions that the number of a priori goals or requirements depends on project size. One interviewee is also adamant that there should be different levels of requirements depending on the project status, for example, ideate, define, or explore, and that the expectations should be aligned on each status segment.

It is not uncommon to formulate goals that just state "improvement compared to current" in terms of some business parameter. Sometimes the parameter to improve is not even defined. Statistical or model-related goals are rarely presented to the stakeholders, unless there is a lack of business metrics to present. They are, however, measured by data scientists but can mean vastly different things. For example, the same accuracy can be either terrible or terrific, depending on the domain, problem, and goal. A domain expert mentioned that when measuring for example forecast improvement, the measure has to be carefully selected based on the domain values, and even though goals are mostly set as improvements to the current state, they can be measured in many different ways.

Most interviewees stated that in the end, business values should drive both goals and evaluation. These goals can come from higher up in the organization, from the team, or from stakeholders, and when goals or requirements are set within the team, the different roles can provide different expertise. Interviewees consider it beneficial to have one role define the domain and business value, and another define the technical aspects. Further, it might be hard for stakeholders to set goals, especially when they are new to analytics, which can be handled by closer cooperation with AdA. AdA can suggest targets and then goals and requirements might develop as the collective understanding is improved around the problem.

When it comes to evaluation, interviewees mention that it is much easier to evaluate specific and defined goals, but that these are harder to formulate. Interviews and observations show that AdA is not always setting specific goals, but rather testing and evaluating ad hoc during the project. Examples of thorough evaluations are lacking in general, but one interviewee mentioned an example of thorough evaluation through simulation, where the model could be evaluated on predictive ability and sensitivity.

Apart from the product success evaluation, there are mentions of learning success in the sense that the main goal of an exploration can be a learning or competence-building experience. This is considered more relevant if the customer organization is less mature in terms of data and analytics. Management interviewees clearly focus more on this parameter, connecting it to organizational transformation. It could also be valuable to set goals for how to build competencies out of a scrapped project. This could provide value in future explorations, according to one interviewee.

4.1.3 Decision making

Currently, the formal exploration process involves multiple decision points on whether to continue or not, but both interviews and observations show that the conditions and criteria for evaluation are unclear. The development team often has to define success themselves, and the evaluation of success feasibility is subjective. According to the official process, the process owner is responsible for coordinating a forum where all stakeholders gather to make decisions on whether explorations have met their goals. They can then obtain a sign-off from stakeholders, indicating that it is ready to be used. However, interviewees mentioned that in practice some people follow this process while others do not. Often, the decision is made by the most engaged person, without a formal sign-off. There are also cases where the team directly starts applying the outcome of an exploration if it shows clear improvements, without going through the formal process. The decision is then shared with the organization afterward.

It is also mentioned that AdA may sometimes be too quick to accept projects without thoroughly investigating their feasibility. There are also challenges in closing started projects. There are examples of explorations that have been ongoing for years without perceived progress, with one explanation being a loss of sponsorship and support from the organization. Interviewees highlight that a reason for both these problems could be a lack of using tangible requirements for decision-making. For instance, if an exploration does not pass an initial data accessibility evaluation, it should not continue.

4.1.4 Cooperation with stakeholders

Many interviewees stated that AdA is not like a typical IT organization when it comes to maintaining and monitoring solutions, meaning that they are not considering future ownership, maintenance, and operations as a part of their scope. That

competence is needed in the receiver organization after delivery. However, this view is not entirely shared by some stakeholders, as it is mentioned that there is a culture of request and delivery from their side. AdA is then seen as a support function, to which a problem can be handed and then receive a solution, and customers do not always realize that they are expected to participate. Further, they do not expect to handle or maintain any automated models, which is thought to be the role of an IT department. This could be due to a general lack of knowledge of data and analytics, or due to cultural aspects.

The technical know-how of the stakeholders usually determines who is leading or driving an exploration, as well as the overall stakeholder involvement. In some cases, stakeholders are involved in the development, and in some cases, they are not. In the latter case, it's more of a request, follow-up, and delivery process, but in most cases, however, there is at least collaboration around defining the current state and the problem. Interviewees also mentioned geographic and organizational location as factors that improve involvement. It can also depend on the project, the problem definition, or whether stakeholder involvement is perceived to enhance value. Request-delivery processes might decrease the possibility for the customer to receive and use the solution, due to lacking technical know-how and low understanding of the product. Though the project managers try to inform about the issue, interviewees find that projects rarely stop due to low involvement. Instead, they are deployed and maintained by AdA themselves in order to create value. This results in deviations from the perceived purpose of the AdA team, but even when the solution is maintained by the customer, Data Scientists get programming questions or act as a support team.

4.1.5 Documentation

The teams have a formal structure and guidelines for documentation, which includes the use of a DevOps board with a card for each project, a Confluence page (a collaborative documentation tool) where more extensive documentation is added, and demo sessions to share knowledge. It is however noted that the content in the documentation may vary depending on the writer, as different roles may focus on different aspects. It is also mentioned that high-quality documentation is more likely to be produced when the goal is to develop a specific product, compared to more exploratory work. Additionally, if an exploration only results in learning, the documentation may be of lower quality or less extensive. All resulting in inconsistency with the above-mentioned structure and guidelines. It is however mentioned that the addition of the role "Exploration lead" has added some level of structure to all projects, but without removing too much freedom.

Interviewees highlighted the issue that there is not always a specified owner of documents and artifacts, and then no one is responsible for ensuring that they are updated and revised. As a result, information from previous projects can be hard to find, leading to duplication of work. The evaluation of "way of working" at all levels is not done in a structured manner, and this is identified as an area of improvement.

According to the internal process description, a standardized Quick Evaluation should take place when defining an exploration to identify challenges that need to be addressed before development. There are also value estimations, called ballpark value, which is an estimated value addition or cost saving created by implementing the suggested solution. However, it was mentioned by one interviewee that following up on the actual value created by explorations is not well established. One domain expert mentioned that ballpark value may not add value for themselves, but it might be valuable for others. Adding to the documentation procedures, the explorations also have a Definition of Done (DoD), where a high-level evaluation checklist is posted. It is, however, the same for all explorations.

When analyzing the documentation of past explorations, it was found that compliance with the aforementioned practices was lacking over the documented period (2020-10-30 to 2023-04-11). Only 34.1% of explorations utilized the quick evaluation practice and 71.85 % had a ballpark value. It should be noted that the quick evaluation compliance has decreased over time, from 76,9% for explorations created in Q1 2021 to 2,27% in Q1 2023.

4.2 Themes

The interview coding resulted in 510 text snippets and 35 unique codes. Figure 4.1 shows the distribution of unique mentions of the challenge codes, and Figure 4.2 shows the success factors. These graphs could indicate which factors and challenges are most present. A short summary of the code snippets in each code can be found in Appendix D. Some codes relate to a specific theme but during the thematic analysis workshop, presented in subsection 3.5.1, it was found that several of the codes contained nuances that paired with multiple themes. Table 4.1 shows the relations between the codes and themes, as well as indirect influences and aforementioned nuances of the statements in a code that relates to another theme.

The coding of interviews, together with observations, have been grouped and reduced into seven main themes. These were then discussed in the focus group, and those results are presented in the end of each theme section.

Note that some interview quotes have been translated from Swedish to English. However, it will not be disclosed which quotes are translated, in order to secure anonymity.

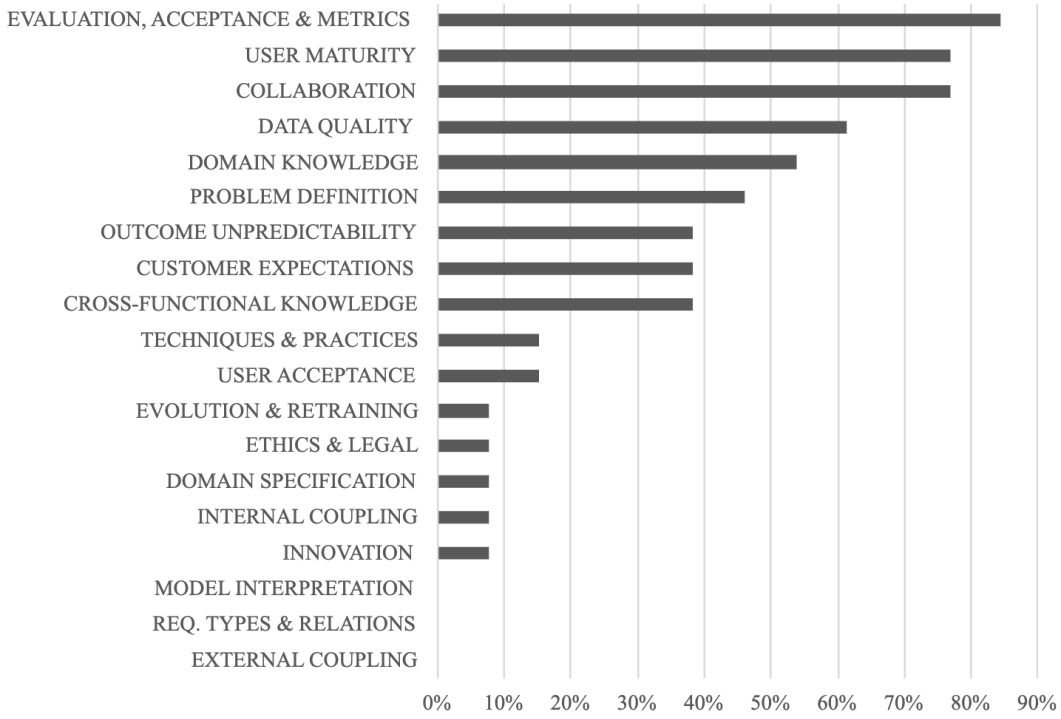


Figure 4.1: Distribution of distinct mentions of challenge codes in interviews.

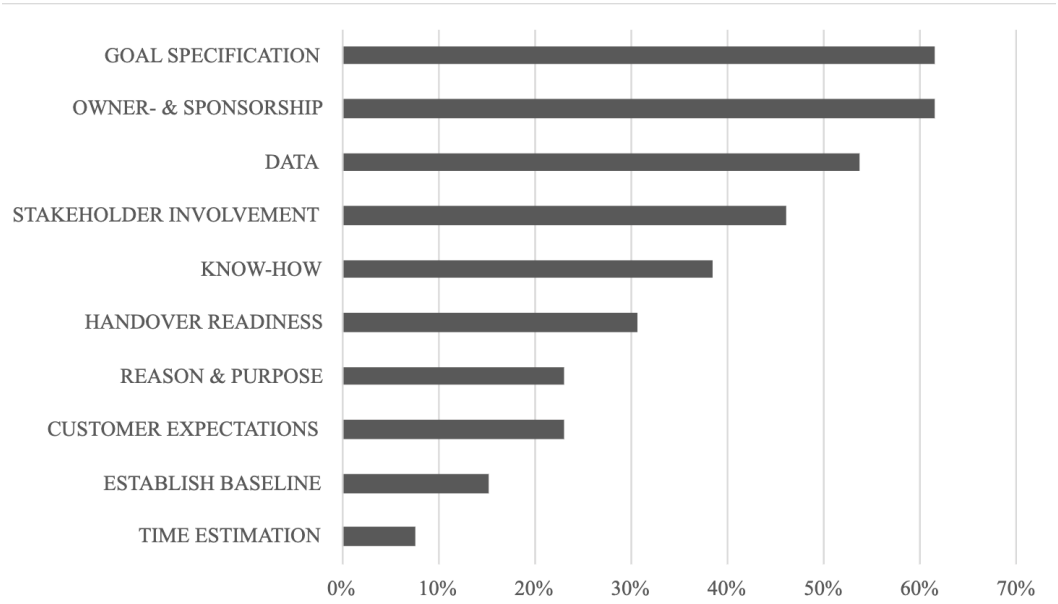


Figure 4.2: Distribution of distinct mentions of success factor codes in interviews.

Table 4.1: Relations between codes and analytical themes. Codes with nuances or indirect influences in other themes are shown in parentheses.

Theme	**Related challenge codes**	**Related success factor codes**
Data & Domain	Data Quality Domain Knowledge Domain Specification Internal Coupling Evolution & Retraining	Data
Ethics & Legal	Ethics & Legal	
Stakeholder Involvement & Cooperation	(Collaboration) (Customer Expectations) Innovation User Acceptance	Customer Expectations (Handover Readiness) Stakeholder Involvement
Structure & Resources	Collaboration (Cross-Functional Knowledge) (Data Quality) (Evaluation, Acceptance & Metrics) (Outcome Unpredictability) Techniques & Practices (User Maturity)	Owner- & Sponsorship Time Estimation
Goals & Evaluation	Customer Expectations (Domain Knowledge) Evaluation, Acceptance & Metrics Outcome Unpredictability Problem Definition	Establish Baseline Goal Specification
Technical Knowledge	Cross-Functional Knowledge User Maturity	Handover Readiness Know-How
Project Purpose	(Innovation)	Reason & Purpose

4.2.1 Data & Domain

> *"Starting with a lot of available data is the main predictor of success."* - Data Scientist

Knowledge about data and its features is a crucial success factor. However, it can be challenging to achieve, due to the varying domains and the perceived distance between developers and the domain. Further, the quality of said data is of great importance to achieve success. Mentioned quality parameters include consistency, number of data points, and reliability, but the quality is mostly mentioned in general terms and can include different aspects for different projects.

> *"If you have multiple people working on the same thing, for example for different brands or for different region or for different scopes, but they do the same thing, then sometimes they do things slightly or more differently. So then, if the data that you're gonna use in the machine learning model comes from decisions that people make, if their decisions are not made in the same way or saved in the same way in in terms of how the data is structured right then you can't use all the data."* - Data Scientist

Both current and past projects have faced challenges in obtaining access to the right data and ensuring its quality.

> *"[Data] is something that often becomes a limitation. It is also common to fail because the quality is not that high. You don't have that many observations (...). We try to use advanced models, but there's not that much to train the models with, and the results will be accordingly [worse]."* - Domain Expert

Given the right data, it is also important to specify the domain in which the product should operate and the characteristics and features of the data. It could further be valuable to predict the rate of change in the domain, in order to estimate the required adaptability and generalizability of the model.

In Figure 4.3 the result from the questionnaire is displayed. As shown the respon-

ID	Statement	Mode	Median	Distribution	No opinion
1.0	**Data & Domain**				
1.1	Data quality must be assured before starting an exploration	2	2,5		0
1.2	Achieving data quality is a major challenge	4	4		0
1.3	Data is unusable without domain knowledge	2 3	3		0
1.4	Changing or inconsistent domains is a challenge in explorations	3	2,5		3

Figure 4.3: Questionnaire results for the theme Data & Domain. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree & 4 = Strongly Agree.

dents agree that achieving data quality is a challenge, but not that it has to be assured before starting an exploration. During the following discussion, the need to ensure data quality was found to depend on the goal of the exploration, and it was suggested that an exploration could focus solely on finding and establishing data quality. If the goal is to develop a model or to solve a specific problem, then low initial data quality increases the risk of not being able to achieve that goal. It was also highlighted that if that is the case, then it needs to be communicated to the stakeholders so that they have the right expectations.

There was more disagreement among the respondents if an inconsistent domain was a challenge or not, but it was largely due to different interpretations of how to define a changing domain. One participant also noted that if it happened it would be a huge challenge, but he had never experienced it in this team.

4.2.2 Ethics & Legal

Societal factors, which are often described as soft factors, were only mentioned by one interviewee. In this case, it concerned the possible legal aspects to consider when deploying an ML model in a specific context, and that there might be limitations to what decisions the models could be allowed to make in a fully automated system.

> *"In practice [the model] is very usable for [the customer] as a recommendation system, but they wouldn't let it take decisions so on itself, because there are legal issues that can arise if it takes the decision."* - Data Scientist

As shown in Figure 4.4, the respondents generally disagree that the projects have ethical or legal implications, but on the other hand, agree that it is challenging to handle those aspects, should they arise. Further confirmation was given when the focus group stated that ethical or legal aspects are usually not as relevant. This is likely due to the business area, logistics for spare parts, but it could be highly relevant if they had projects with HR or Customs, for example. Note the number of respondents with no opinion on these statements, which further indicates a lack of consideration in general for these questions.

ID	Statement	Mode	Median	Distribution	No opinion
2.0	**Ethics & Legal**				
2.1	Models or products have major legal or ethical implications	2	2		4
2.2	It is challenging to manage legal and ethical aspects	2 3	3		3

Figure 4.4: Questionnaire results for the theme Ethics & Legal. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree & 4 = Strongly Agree.

4.2.3 Stakeholder Involvement & Cooperation

Involvement and cooperation with the stakeholders and customers are mentioned as contributing factors to the success of a project, primarily in two ways. Firstly, it is necessary to gain their insight to develop the right thing, and to develop something that brings actual value, as they have the necessary domain knowledge.

> *"[The requirements] is something that I might have an idea about, but I'm not the one that has to use the result. It is something that definitely has to come from the domain experts."* - Data Scientist

Secondly, if they are involved they get the right expectations on the outcome of the project, and what is expected of them in terms of technical competence to be able to use and maintain the final product.

> *"(..) but that the stakeholder is ready enough to understand that 'we need to participate and learn through the entire exploration. Not the domain, but the technical. So that we are ready to take care of this product'."* - Data Scientist

This however comes with challenges. The involvement requires interest, and that depends on the stakeholders' understanding of the projects' value. This suggests that the focus should be on solving their existing problems, but this could limit the level of innovation. The solution that might be the most innovative and have the largest impact might be difficult to grasp and directly connect to the current list of identified problems.

> *"Many ideas arise from the will to test a technical solution. Then you test something, and it might be super cool. But I think those [explorations] risk becoming a bit disconnected. If it happens to be in a segment where 'well, what you're trying to improve already works fine for us, but over here it is critical right now. Why aren't you helping us with that instead?'. That can create a bit of a gap, or at least a percieved gap."* - Domain Expert

Another challenge relates to a culture of request and delivery processes. The customer might expect to simply hand over a problem and then get a solution ready to use, which is not optimal according to the interviewees. Interviewees mention successful examples of projects where one person has embodied all aforementioned roles and hence was solely cross-functional. While the projects were successful, interviewees agree that this is neither a scalable, nor optimal, solution to the cooperation problem.

The importance of stakeholder involvement was further confirmed in the questionnaire, as presented in Figure 4.5. Both the mode and mean of these results show that respondents were in relatively strong agreement with all three statements. In the following focus group, the definition of a stakeholder was briefly discussed. Some

ID	Statement	Mode	Median	Distribution	No opinion
3.0	**Stakeholder Involvement & Cooperation**				
3.1	Stakeholder involvement is necessary to achieve exploration success	4	4		0
3.2	Involving and cooperating with stakeholders is challenging	3 4	3		1

Figure 4.5: Questionnaire results for the theme Stakeholder Involvement and Cooperation. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree & 4 = Strongly Agree.

participants stated that all beneficiaries in a project should be considered stakeholders, for example, both data scientists who build competence and a customer that receives a product. On the other hand, some viewed the end user as the stakeholder to be considered.

The focus group discussion also diverted into a discussion around how to define exploration success, and how that definition could influence the importance of stakeholder involvement. Similar to the interview results, it was stated that if success is defined as the adoption and usage of a product, then stakeholder involvement from an early stage is crucial. If, however, the goal is to learn and discover technologies, stakeholder involvement has less impact.

Finally, the focus group mentioned that a continuous feedback loop is always important, no matter the stakeholder definition.

4.2.4 Structure & Resources

Firstly, the ownership and sponsorship of explorative ML projects are important to the interviewees and are approached both as a success factor and a current challenge. The owner has to keep the project on track and push forward, in order to keep the interest of the stakeholders. However, it was mentioned that developers tend to haste into the building of the models and that challenges arise when there is no data available, or if the scoping and defining of the problem are lacking.

> *"I think that something essential to succeed is that you have someone that is a good sponsor, someone that owns the problem. What we do is often based on a problem somewhere, or a possible solution or automatization. I think, no matter the project, that to have a clear ownership, that is the most important part."* - Manager

Secondly, resource and time allocation are challenging when working on cross-functional projects. The data scientists have a clear sign-off to spend time on explorations, but involved stakeholders might be unable to allocate and prioritize the same amount of resources. The cross-functional team can also be a challenge or success factor since involvement is beneficial for understanding, competence, and learning, but a larger team might decrease efficiency. The decrease in efficiency is due to more time being spent on administrative tasks, according to one interviewee.

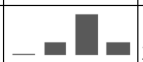
ID	Statement	Mode	Median	Distribution	No opinion
4.0	**Structure & Resources**				
4.1	A defined owner or sponsor is necessary to achieve exploration success.	2 3 4	3		1
4.2	Implementing the model in parallel to defining the data and targets generates more challenges with data quality or problem	3	3		2
4.3	Involving non-technical actors in an exploration decreases development efficiency	1	1		2

Figure 4.6: Questionnaire results for the theme Structure & Resources. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree & 4 = Strongly Agree.

Some interviewees also mention a lack of understanding of the R&D-like nature of explorative ML projects.

> *"I would say that ML is more close to R&D on a scale (...) and I think that some people have a problem understanding that. What we are really doing is operations analysis; analysis projects to take better decisions. These kinds of projects does not really fit into the agile or Scrum approach, because it is more explorative and inclined towards research, rather than projects and implementations (...). I've noticed that it's a challenge to get people to understand that it is not possible to produce analytics in a streamlined manner, only using agile and Scrum . At least not right now. It might be possible when we've learned more (...) and found a better method to do it."* - Domain Expert

Lastly, some interviewees argue against front-loading requirements early in the process and suggest a more incremental approach. There are, however, notions of the trade-off between freedom and knowledge of potential pitfalls, as mentioned in the first paragraph.

The result from the questionnaire can be seen in Figure 4.6. There was no unified opinion on whether a defined sponsor or owner was needed. One participant mentioned that general sponsorship is needed, but it does not have to be a specific person. There was further some consensus that a sponsor or owner is not needed in the early, more exploratory stages. However, it could be beneficial when moving towards a product, to facilitate the decision-making. Without it, there is a risk that a model with unclear results gets stuck in limbo where it is not scrapped but not implemented either.

Further, the results show that involving non-technical actors does not necessarily decrease development efficiency. In the discussion, it was mentioned that in the actual model development stage it might decrease the efficiency, but with the general exploration, efficiency might instead increase as they could provide other non-technical input, for example, business aspects or domain knowledge.

4.2.5 Goals & Evaluation

The challenge of evaluating model results and determining success is present throughout almost all interviews. Note the spike in the mentions of the code *Evaluation, Acceptance & Metrics* in Figure 4.1. It could either be a matter of not having enough well-defined goals to evaluate against, or it can depend on the lack of knowledge of the current baseline. As an example, the forecasting accuracy of an ML model can be considered arbitrary. For some contexts, even a low accuracy can provide value, while other applications demand very high scores.

> *"Yeah, [evaluation] is a bit arbitrary. That's why we like to look at business value at the end of the day. Because we have seen, for instance, when we when we created the model, we got some number that was not really best, something like 80% accuracy. Which is not bad, but it's not something to (...) base how we know if the project really helps. So, should we implement the model or not? And this, I think it's arbitrary. So we were surprised to find that even if the model is not so spectacularly accurate it's still quite usable, (...) it creates business value."* - Data Scientist

Further, evaluating improvements can only be done if a measure of current performance is established. Without it, there is nothing to compare against. Not having a clearly formulated goal, or not documenting the goals, can also make the evaluation more arbitrary. Goals can then be gradually altered over the course of the project, making the end evaluation disconnected from the initial problem or goals.

> *"A common thing is that you should have a idea of what you want to do prior to starting. I would call that a critical success factor."* - Domain Expert

Setting targets for a project is also strongly connected to the uncertainty of the outcome. Interviewees find challenges both in estimating project value and in managing expectations from stakeholders since it is hard to guarantee any specific outcomes. Expectation management, as well as goal specification, is also connected to the definition of the problem at hand. Knowing the customer, the problem, and envisioning the end goal of the project is challenging but highly important in order to formulate specific and measurable goals together with the stakeholder.

> *"And to make sure that we don't only set specific goals for the project, but also have a strategy on where we're going to go with it."* - Data Scientist

Lastly, interviewees mention that goal-setting and evaluation tend to look different depending on whether the target is to deliver value to a specific stakeholder or to explore and learn new technology.

The focus group discussed that setting specific goals can be very challenging. A broader formulated goal, such as "we should improve this process", is easier, but to define specific, measurable, targets is harder. Further, the focus group concurs with

ID	Statement	Mode	Median	Distribution	No opinion
5.0	**Goals**				
5.1	A clear problem statement is necessary to achieve exploration success	4	3,5		0
5.2	It is challenging to define specific goals early in an exploration	3	2,5		0
5.3	It is challenging to establish a baseline or current state to compare against	3	3		1
5,4	Requirements and specific goals hinder creativity in explorations	3	2,5		1
6.0	**Evaluation**				
6.1	It is challenging to evaluate exploration success	3	3		1
6.2	Evaluation is improved by having specific goal- or target parameters	4	3		0
6.3	Explorations should be evaluated mainly by stakeholders	2	2		0
6.4	Domain knowledge is necessary to evaluate models	3	3		0

Figure 4.7: Questionnaire results for the theme Goals & Evaluation. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree & 4 = Strongly Agree.

the interviews in that finding the baseline or current state is highly challenging, which is also shown in the questionnaire, Figure 4.7. The focus group mentions that a lacking specification of goals and targets might simply be a consequence of the difficulties with it.

When discussing evaluation, the focus group also agreed with previous results, stating that when delivering a product to the stakeholder, they have to be highly involved in the evaluation. However, for technological explorations, the evaluation of learning experiences can mainly be done by the exploration team.

The focus group agrees that specific goals improve the evaluation process, as shown in Figure 4.7, but that it can be challenging depending on the exploration. The level of specificity can also be a matter of exploration maturity, as an early exploration on a topic might be less specific than a subsequent exploration that is closer to an implementable product.

Lastly, after discussing the importance of specific goals, the focus group discussed whether there is a trade-off between creativity and requirements. The participants found it a bit ambiguous, as reflected in Figure 4.7 as well, as a project with no boundaries can be too open to be creative, but an overly constrained project limits freedom. They thought that there is a sweet spot between the two, that might be different from project to project. Some projects where previous components are reused or repurposed might benefit from specific requirements and a streamlined process, similar to Mode 1 rather than Mode 2. Other projects need to be more open, in order to find an innovative solution.

4.2.6 Technical Knowledge

Technical knowledge is mentioned as both a factor of success and a challenge. The technical knowledge can be split up into two parts. The first one is knowledge within the development team and the second one is the competence of the customer. The first one is mentioned as important but is not necessarily a problem in this case, as the technical competence within the team is generally high. The important aspect is to make sure that it is available in all projects from the start, and then continuously throughout the projects.

The second one poses a big challenge. The customers' technical knowledge is crucial to be able to formulate the problem and set goals, to evaluate if the product has met said goals, and to be able to use and maintain the final product. It is a balance between making the product easy to use and raising the competence of the customer.

> *"In the best case scenario I think that [the requirements] should come from those who know the business best. It should be like 'I know how Volvo creates value for the customers through my process. That process could create more value if we had some AdA in it'. That requires a quite mature business organization, which [currently] might not be that used to handling ML. But when it works perfectly, I think it is built in to our improvement reasoning that 'I know there is potential to add some ML in this process'."* - Manager

> *"It's always easier (...) that [the customer] have the technical competencies to take over [the product], or at least be able to run it and use it"* - Data Scientist

The responses to the questionnaire statements concerning technical knowledge can be seen in Figure 4.8. Regarding the technical knowledge of the stakeholders, the respondents agreed more with it being necessary to use the final product, rather than to collaborate. It was mentioned that when collaborating, the more technically competent actors could compensate for the lack of technical knowledge of the stakeholders, and stakeholders could instead provide other valuable input. It was also stated that the product can, and should, be adapted to the technical knowledge of the user and therefore technical knowledge might not be as critical. However,

ID	Statement	Mode	Median	Distribution	No opinion
7.0	**Technical Knowledge**				
7.1	Exploration teams have a sufficient competence mix	3	3		2
7.2	Technical knowledge of the stakeholder is necessary in order to collaborate	1	1		0
7.3	Technical knowledge of the stakeholder is necessary in order to use the product	2	2		0

Figure 4.8: Questionnaire results for the theme Technical Knowledge. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree & 4 = Strongly Agree.

there was consensus that technical knowledge improves collaboration, but not that it is a necessity.

Another mentioned aspect was that the user has to trust the model or product in order for them to actually use it, both trust the output and be confident in using it. This trust was mentioned as easier to establish if the user has technical knowledge."

4.2.7 Project Purpose

Interviewees mention that there needs to be a valid reason behind the initiation and continuation of an exploration. Some state that it is always important to have a business context, and to avoid doing things just because of a wish to apply a new technique or solution. In contrast, other interviewees mentioned the importance of learning, and that it can be limiting to only consider the business value and not other aspects that can be gained from a project. But if you are to find disruptors, one interviewee argued, you cannot just work on what you already know.

> *"If I want to put a nail in a wall, I use a hammer and if I want to put a screw in the wall I use a screwdriver. Not state 'I want to use a hammer' and then there is a screw. It works, but it's not the right tool. People also say 'I want to use the hammer, now I need to find a wall to put a nail in'. Then you might think 'this wall might look better with a nail', but is it necessary?"* - Data Scientist

ID	Statement	Mode	Median	Distribution	No opinion
8.0	**Project Purpose**				
8.1	There should always be a specified connection to business value in explorations	1	2		0
8.2	Technological explorations should not be evaluated on their business value impact	3	3		0

Figure 4.9: Questionnaire results for the theme Project Purpose. 1 = Strongly Disagree, 2 = Disagree, 3 = Agree & 4 = Strongly Agree.

In contrast to the differentiated perspectives in the interviews, the focus group was almost in consensus that there can be explorations without specified business value. For example, there can be entirely technical explorations, focused solely on learning. It was also stated that the indirect value of learning should not be forgotten, for example through events and activities that promote the organization's knowledge about analytics.

The questionnaire results, Figure 4.9, show that respondents are aligned on that technological explorations should not be evaluated on their business value impact. The responses for question 8.1 was slightly more diverse, but still generally disagrees with the need for specific business value in all explorations.

5
Discussion

The purpose of the study was to provide empirical case evidence to contribute to the existing body of knowledge on RE, agile RE, and RE for ML. The purpose was further to identify suitable RE practices in other fields of SE, and suggest them for further validation in the field of ML development. This chapter will discuss the case study result in relation to the literature presented in chapter 2, divided into the analytical themes found in the study. The discussion is then summarized and connected to the research questions. Lastly, this chapter contains a discussion regarding threats to validity and a summarized suggestion for further research.

To navigate the discussion, Table 5.1 depicts the themes and key takeaways.

5.1 Data & Domain

Data quality is central to ML project success. Data Requirements should be separated, and a certain data quality assurance should take place before initiating projects, to pass or fail them without investing excessive resources. Stakeholders should be involved, in order to understand and specify the domain. The Pyramid Model [31] is suggested as a practice to facilitate ordering of requirements, with Data Requirements at the base of the pyramid.

It is clear from both background and results that data has a critical impact on both the development process and the outcome of ML products. This has to be approached from different perspectives, taking consistency, security, variety, and other quality aspects into account. This is in agreement with, for example, Amershi et al. [20] and Wan et al. [21], and can be related to the challenges found in the group *Data Dependence* in Table 2.1, and more specifically *Data Quality Issues*. In terms of requirements, it was found that some level of data requirements should be prioritized, as a pass or fail requirement, before starting development. The data does not have to be packaged and prepared before development, but the required level of quality has to be investigated and assured to avoid wasting time in development. The result from the focus group disagrees with this to some extent, with the argument that lacking data quality simply shifts the goal of the exploration to establishing data quality. However, this might be unique to the case context, as it might not always be possible to be flexible with the goals of a project.

5.1 Data & Domain	
Findings	Challenges with data quality assurance, new requirement types, and domain knowledge. Need to improve stakeholder involvement, and prioritize data early on. This study concurs with previous papers.
Practices	The Pyramid Model
5.2 Ethics & Legal	
Findings	Ethical and legal factors were not as present in this case as they are in previous literature, suggesting that the context is important to consider.
Practices	No suggested practices
5.3 Stakeholder Involvement & Cooperation	
Findings	The uncertainty of ML projects require continuous involvement to elicit requirements incrementally, and to manage expectations. User acceptance is also important, being eased by involvement.
Practices	Scenario based requirements elicitation.
5.4 Structure & Resources	
Findings	Being similar to R&D processes, ML development requires sponsorship and clear resource prioritization. Selecting the right project planning strategy is crucial.
Practices	Differentiate clearer between Mode 1 and Mode 2.
5.5 Goals & Evaluation	
Findings	Evaluating project success is one of the major challenges in ML development, agreeing with literature on that its arbitrary measures are harder to specify.
Practices	ML-GORE, Use Case Diagrams
5.6 Technical Knowledge	
Findings	The study indicates that RE for ML, and ML development in general, requires that the customer has high technical knowledge, to formulate requirements, secure handovers and increase acceptance.
Practices	Scenario-based requirements elicitation
5.7 Project Purpose	
Findings	The purposes of projects should be clearly stated, to adapt the requirements process. Aligning these purposes is highly important to balance resources and manage expectations.
Practices	No suggested practices
5.8 Non-Mentioned Theoretical Challenges	
Findings	Coupling, team-internal collaboration, and requirements trade-offs are challenges from literature that were not present in this study.

Table 5.1: Summary of key findings and practices in the discussion. Refer to each section for evaluated discussions.

In this study, formulated data requirements are lacking, and data investigation is integrated with development as a part of exploring what can be done with the data at hand. The results indicate that much time can be saved if data quality is instead, to some degree, assured before the decision to start an exploration. This can increase the chance that the desired product is feasible and meets expectations. As was mentioned in the focus group, lacking data quality increases the risk of the project. Some of the mentioned challenges with uncertainty are highly dependent on the data quality uncertainty, which can be mitigated, or at least identified, through early data evaluation. It is also suggested that it is much more time-consuming to fix bad data afterward, or during development, which further motivates separating data requirements and prioritizing them early on. This suggestion to separate data requirements from other requirements is supported by literature, for example, in the article by Vogelsang and Borg [25], who also argues for this category of requirements for ML.

In order to achieve a proper data requirements process, the results suggest that increased stakeholder involvement could improve understanding of the domain and data features, which was brought up as both a challenge and a success factor. This challenge is in line with findings by Wan et al. [21] and Belani et al. [23], and can be related to the challenge *Domain Knowledge Dependence* in Table 2.1. Specifying the domain together with the stakeholders can reveal what parameters are critical, and what quality issues or requirements have to be prioritized. Continuous stakeholder involvement could also increase the inherent domain knowledge of the technical experts, improving understanding in the long term. As a part of the requirements process, the question of how long the model remains valid also has to be addressed. If the model has a high risk of being invalidated after a short period of time, due to for example a volatile domain or data set, requirements on retraining and model resilience have to be prioritized. This relates to both *Evolution & Retraining* and *Domain Specific Viability* in Table 2.1. It was confirmed to be a potential challenge by the focus group, but it was then mentioned in a theoretical sense, as changing domains was not something the participants faced in their development. This theoretical discussion is supported by Heyn et al. [24] who argue that if the context or domain should change, the output of the model might be irrelevant.

Even though the participants stress that creativity is central, the results and literature indicate that there has to be a more rigid process when ensuring data quality. The creativity might in fact be hindered by a low-quality data process, since more time is spent on finding or refining data, instead of developing and exploring models.

The key aspect of this theme is that data is central and has to be taken into account from the start. The Quick Evaluation (Table 1.1), which is already implemented today but has low compliance, could facilitate this. Its purpose is to estimate different preconditions for a project early on, for example, different aspects of data quality, so that the right expectations and focus can be set. It also helps with assessing the viability of a project and could even provide reasons to scrap a project in the idea phase.

A model that could be relevant to handle these challenges is the Pyramid Model presented by Giunchiglia et al. [31]. The aim of the model is to prioritize requirements in ML development, with one significant aspect being setting and validating data requirements early. As mentioned earlier, fixing bad data later in development can both be time-consuming and costly, and this could be mitigated by the Pyramid Model. It could also be a way of identifying and specifying the risk of loss in performance over time and how it should be handled.

5.2 Ethics & Legal

Ethical and legal factors were not as present in this case as they are in previous literature. This suggest that the challenges related to this are context dependent. The context should be investigated in an early stage, to identify ethical or legal implications to consider during development.

Societal factors like ethical or legal did not have any noticeable implications in this case. It was barely mentioned in the interviews and in the focus group there was only a brief discussion on the topic. The fact that the team works on internal products could affect the occurrence of this. Since they rarely handle personal or sensitive information or act as decision-makers in a legally constrained context. This differs from the literature, which highlights this type of challenge, see *Ethics & Legal Impact* in Table 2.1, as highly important and challenging. For example, Arpteg et al. [19] state that data privacy poses challenges, but this study shows that it is highly relative to the context in which the models operate and what data is used. The example cases in Arpteg. et al. are more inclined towards, for example, user demographics, which makes the data more sensitive.

One should also consider that the interviewees had mostly technical roles, and might not consider the ethical and legal aspects as a part of their workload. Interviewing representatives from a legal- or HR department might have shown different results, but there is little evidence to support it in this case. Since ethical and legal aspects depend on the context, it is important to identify their potential implications in the RE phase.

5.3 Stakeholder Involvement & Cooperation

While stakeholder involvement is not a requirement in itself, it has a positive impact on the requirements process. The uncertainty of ML projects requires continuous involvement to elicit requirements incrementally, and to manage expectations. Practicing scenario-based requirements, for example, to involve stakeholders, could also improve user acceptance.

Some degree of involvement from the stakeholder is said to be crucial for the success of the development. Stakeholder cooperation might not be formulated as a requirement per se, but it is important, especially in the early stages, to be able to formulate the problem and to ensure a proper requirements process. In this case, involvement in the ideation and empathy phases, or any kind of requirements elicitation and specification stage, could have obvious benefits. It is also highlighted that continuous involvement in the development process is preferred. This as the non-determinism and uncertainty of ML projects make it much harder to maintain a request/delivery organization. Greater stakeholder collaboration is clearly needed in order to incrementally increase requirements and understanding of what the finished product should look like.

Cooperation is also a challenge mentioned in the literature relating to *Collaboration Difficulties* in Table 2.1. Amershi et al.[20] and Arpteg et al.[19] however, focus more on the interactions between software engineers and data scientists, and the implications of their different backgrounds and technical knowledge. The fact that this study has focused on the early stages of development and that there is no software engineer in the local team could explain why this aspect of collaboration was not present in this case. Instead, there was more focus on the collaboration with the customer and domain expert, which are roles that are more involved in the parts of development included in the scope of this study. This type of collaboration is present in a paper by Inyat et al., discussing success factors in agile RE[15].

Another aspect that was mentioned by Heyn et al. [24], and strengthened by this case, was the challenge of getting the users to accept the output of the model, relating to *Achieving User Acceptance* in Table 2.1. This could, for example, be a model that replaces a decision-making process that had previously been done out of experience and gut feeling. The implementation of such a model could then be met with a great deal of skepticism. The reason behind this could be a lack of trust or understanding, both of which could be increased with more involvement throughout the development process. Involving stakeholders when formulating scenario-based requirements could, as a formal practice, ease understanding of the usage and relation to the models and products. This could in turn increase acceptance.

The difference between problem-defined processes and technology R&D processes is also something that could be more clearly differentiated. It seems like a problem-

oriented project could, and should, have a higher degree of formal practices, to ensure proper collaboration and specification of for example the problem, domain, and data.

5.4 Structure & Resources

The many uncertainties in ML development makes it similar to an R&D process. Both a specific sponsor, as well as cultural sponsorship in the organization, are suggested as success factors in ML development. Prioritized resources are needed both from the development team, and the stakeholders, in order to formulate and evaluate requirements continuously, in a cooperative manner.

While the process and structure of a project might not be part of its requirements documentation, the study finds that sufficient organization and structure have a positive effect on the ability to formulate and evaluate requirements.

In terms of creativity, the leaders and the organization itself must sponsor and support explorative processes. For the specific exploration, there was some disagreement in the focus group on whether sponsorship is needed or not. However, the disagreement revolved around the need for a specific sponsor or owner. There was consensus that sponsorship was needed in general both to push the project forward and to take necessary decisions on the viability of a produced model.

The involvement of stakeholders in the exploration process is mentioned as having several benefits that are discussed in the other themes. The involvement, however, is to a great extent dependent on them having time to spend on projects that are not part of their operational work. This is another aspect where sponsorship and support influence the project outcome. If the part of the organization where the stakeholders belong does not sponsor or value R&D projects, the involvement from their side becomes more difficult. This then affects, for example, the requirements specification and evaluation.

The literature found did not specifically mention sponsorship in this context, but this is possibly due to it not being directly connected to RE or ML, which the literature search was delimited to. Neighboring to this topic, however, is the fact that structuring and planning the development of ML projects is difficult due to uncertainties [19], relating to some aspects of the challenge group *Uncertainty* in Table 2.1. This could further affect the resource estimation for a project, which could in turn make cooperation with stakeholders more difficult, as it is harder to determine how much time is required from their side.

Further, it was found that it is difficult to structure ML development. The organization in this case utilizes something they call Exploration Mode Development, which

is largely based on Design Thinking, where most projects are similar to Mode 2 development. This is to handle the uncertain nature of ML and to approach it through an explorative setting, where the focus is inclined toward learning and building competencies. There are however split opinions if this approach is suitable in all cases. It is brought up that ML has to be organized more like R&D development, rather than a structured delivery process. However, some projects could benefit from a more structured approach similar to Mode 1 development, for example an agile release train or Scrum for smaller projects. However, the most suitable approach appears to be largely dependent on both the type of project and the organization's maturity in ML development.

5.5 Goals & Evaluation

Evaluating project success is one of the major challenges in ML development. It appears due to a lack of documented requirements, and a shift in requirement types. Further, uncertainty and arbitrarity in some performance measures increases the need for collaboratively defined goals, baselines, and evaluation parameters. ML-GORE and Use Case Diagrams are suggested practices to align actors on these requirements.

This study suggests that the process of evaluating and determining project success or failure is one of the biggest challenges in ML development. It has been present throughout the case study, and a number of reasons have been indicated during the data collection. One of the main reasons is the lack of properly documented requirements and more specifically, the right type of requirements. Even though the projects are explorative, it is crucial to have a goal statement, as well as relevant parameters and measures to use when evaluating. These have to be documented in order for them to remain consistent throughout the project.

Further, it is clear that RE for ML is different in terms of the requirements used, and specifically the level of determinism in these requirements. ML projects require an understanding of which measures to use, but also of the level of importance and influence of these measures, relating to the challenge *Evaluation, Acceptance & Metrics* in Table 2.1. Accuracy is mentioned as a good example, which can mean vastly different things for different projects. The arbitrary nature of some performance measures is also mentioned by for example Vogelsang and Borg [25], and the importance of establishing a baseline to compare against is clear. Ishikawa and Yoshioka [5] argue for more continuous evaluation with the customer or stakeholder, which this study supports as well. Close collaboration in goal setting, baseline definition, and evaluation is highly suggested to cope with expectations and uncertainty as well.

We find that many interviewees, as well as papers by Wan et al. [21], Giray [18], and Vogelsang and Borg [25], find that lacking understanding of ML technology tends

to lead to expectations of perfect products, which relates to *Customer Expectations* in Table 2.1. When working with goals and targets, a great effort has to be put into the mutual understanding that ML models are statistical and probabilistic, not deterministic products from which one can expect the same result every time. It is once again a question of non-functional requirements, for example, how accurate a model is compared to the baseline, and not functional requirements where a certain input or action generates a certain specific output. These measures of quality aspects are currently missing and should be incorporated in the RE practices of ML projects.

To cope with these challenges, the study suggests the use of Goal Oriented RE for ML (ML-GORE) to better structure the path toward project success. Ishikawa and Matsuno [4] also suggest this evidence-driven RE approach, that connects to GORE, where uncertainties are identified and a hypothesis for each uncertainty is formulated. Tests or experiments then confirm or deny these hypotheses and drives the project forward. It should be noted, however, that ML-GORE is a relatively complex framework, and while it could benefit the organization, it might require a higher maturity in RE before implementation.

To begin highlighting uncertainties, it is suggested that the organization increases compliance on, and visibility in, their quick-evaluation process (Table 1.1). This can set the goals in a context of risk and uncertainty, which together with an evidence-driven approach might increase success rates.

Use case diagrams [14] can also be a simple but effective practice in RE for ML. It could improve requirements elicitation, but can also add to the users understanding of how they interact with the model and what expectation levels are reasonable.

5.6 Technical Knowledge

ML projects put high demands on the customer's technical knowledge. They have to be knowledgeable enough to formulate and evaluate goals and requirements, but knowledge is also required to use and maintain the final product. Comprehension of the model can also decrease if the technical knowledge is insufficient, suggesting that stakeholders should be actively involved and learn throughout the project. Scenario-based requirements elicitation could provide bidirectional information on the stakeholders' knowledge, to adapt the product and increase learning.

The study indicates that RE for ML, and ML development in general, requires that the customer has high technical knowledge. This is approached from two angles. The first one is the effect it has on the process of defining goals and requirements, and then later evaluating if these have been met. As seen in this study, and also mentioned by Wan et al. [21] and relating to *Req. Types and Relations* in Table 2.1,

functional requirements are not as relevant for ML. Instead, the focus is more on quantitative requirements which could be more difficult to formulate and understand without a certain level of technical knowledge.

The second angle is in the delivery phase, where it is mentioned that in order to take over, use, and maintain the product, a certain level of technical understanding is needed. The final product seems to be of a much more technical nature, considering both the output it produces and how to use it. The performance of an ML model might also decrease over time, which is something that has to be monitored and maintained by the customer. That the product user is responsible for the maintenance and monitoring of a product's performance is unusual. In a request and delivery situation, the technical responsibilities of the customer would be less.

Technical knowledge was mentioned by almost all interviewees and, is something that was noticed to be a major focus area of the organization. The problem is almost only approached from the angle of increasing the technical knowledge of the customer; only one interviewee mentioned the option to make the product easier to use. This could be because of a lack of SE thinking within the team, where Data Scientists or similar roles define the level of usability based on what product they want to deliver, not what the customers require. Another factor could be the lack of requirements processes, where the question of usability might have been brought up.

This theme connects to the groups *Multidisciplinarity* and *Comprehension* in Table 2.1. Amershi et al. [20] and Arpteg et al. [19], for example, mention challenges with technical knowledge due to the complexity of ML, but the focus is not specifically on the customer. It is instead discussed in a broader sense that there might be issues with comprehension of the model and the output it produces. Wan et al. [21] also highlight the importance of increased knowledge for requirement engineers to be able to set better requirements. Nascimento et al.[3] further suggest that management should have greater knowledge in RE and ML, which could enable collaboration through sponsorship and resource prioritization. The latter is strongly supported by this study since stakeholders lack time to actively collaborate in a sufficient way.

Another aspect of lower technical knowledge is that it could affect the level of innovation in delivered products. As mentioned in section 5.3, involvement from the customer is almost necessary for success, and as indicated by this study, the involvement is easier to achieve when the customer understands the value of the project. A more innovative approach to the problem could be more difficult for the customer to connect to their specific operations, making it harder to see the value created.

The handover processes might be specific to our context, but user maturity is still highly relevant in a more general sense. Maintenance of the product is also important to consider in the RE phase, and requirements have to be aligned regarding the usability of the final product. Viable prototypes are also more difficult to produce in ML systems, in comparison with for example design prototypes of web pages or

similar applications. This could result in it being more difficult to identify problems with usability early on. Mentions of user maturity for ML in literature are lacking, and we find that research is mostly considering ML as integrated parts of a greater system, where the user of the system does not interact directly with the output of the ML model. However, Heyn et al.[24], for example, mention difficulties in understanding ML decisions or outputs, which is neighboring to user maturity.

Practicing for example scenario-based requirements elicitation could foster the discussion around user maturity, and how the user interacts with the model. Cirqueira et al. [32] also argue that this can increase explainability. When presenting usage scenarios, the developing team could also get valuable information on the technical knowledge of the customer and can adjust expectations on the simplicity of the final product. Scenarios could also be formulated at different levels of detail, increasing throughout the project.

To further assess and more specifically document the technical knowledge of the user, it could be added as a field in the quick evaluation (Table 1.1) practiced by the case team.

5.7 Project Purpose

The purpose of a project should be clearly stated, to adapt the requirements process. A problem-oriented project should have a business value proposition, and should be evaluated accordingly. Technological explorations, on the other hand, should focus on learning outcomes and competence-building. Aligning these purposes is highly important to balance resources and manage expectations.

The results show that there are two general reasons for starting an exploration, either it is based on a problem with a business value proposition, or it is based on technological interest.

Much like innovation and R&D processes, this study suggests that business value should be a driver, in conjunction with the sole value of learning more and being innovative. However, it is suggested that if a project can be clearly defined as problem-oriented, it should have a business value statement that can be used as a basis for evaluation. A technological exploration, on the other hand, could be more focused on the level of learning, and should in turn have fewer expectations to deliver short-term business value. In this way, the organization's expectations of the team and the process can be more realistic, and prioritization can be made whether the team should focus more on one or the other. This can also have an effect on resource allocation decisions, which can be a driver or inhibitor of stakeholder involvement.

This theme does not directly connect to any of the challenges found in the literature. This could be, similar to section 5.4, because the search for articles was delimited to RE and ML. *Project purpose* is a more high-level topic, not directly related to RE, but as previously mentioned it greatly affects other parts of a project. This is especially true in the explorative context, as there could be projects whose sole purpose is to learn.

Lastly, unlike most other themes, the responses on this theme in interviews and the focus group differed slightly. This speaks to the nuances of the value propositions in explorative projects. On the one hand, it might be easier to identify direct and short-term value, but in a discussion setting such as the focus group, it becomes clear that in a long-term strategy, learning and technological evolution has to be considered valuable as well. The balance between the two is important, but out of scope for this thesis. It is however clear that the team needs to be in agreement on what purpose the current project is believed to fulfill.

5.8 Non-Mentioned Theoretical Challenges

Coupling, team-internal collaboration, and requirements trade-offs are challenges from literature that were not present in this study. This might be impacted by the organizational maturity, or its role composition, as well as the scope of the study.

Although this study finds evidence that agrees with most previous literature, there are some cases where the study provides little evidence for, or against, theoretical challenges.

Firstly, the challenges in the group *Coupling* in Table 2.1 were not present in this case. External coupling was not mentioned at all, and internal coupling was only briefly mentioned by one interviewee. Many articles mention coupling as an ML challenge ([20] [19] [23] [26] [21]), but the context of this study might disguise the impact in this case. For example, ML models in this case are less integrated in a larger system, which decreases the impact of external coupling. The organization is also less mature in terms of quality assurance of code, which might suggest a lower awareness of internal coupling.

Secondly, as mentioned in section 5.3, this study finds challenges in collaboration with the customer or user rather than internally in the team, which is mentioned by Amershi et al. [20] and Arpteg et al. [19]. This study might not show internal collaboration difficulties since it focuses on the early exploration processes rather than processes for scaling out and maintaining products. For example, these early stages does not, in this case, involve software engineers, that could introduce differences in culture, processes, or previous experience.

Lastly, the study finds that new types of requirements are present, but does not find that the case considers the new relations and trade-offs between requirements mentioned by Horkoff [26]. The lack of data on that topic most likely depends on that the organization currently lacks experience with RE practices and might not identify or acknowledge these challenges.

5.9 Summary and Research Questions

To summarize this chapter, the discussion around, and answers to, the four research questions are presented below.

5.9.1 RQ1: Which are the major observable challenges present in RE for ML processes?

The study finds that RE for ML differs from regular RE in many ways. The dependence on data makes data quality requirements critical, and the uncertainty of the outcome impacts the RE processes regarding both expectation management and stakeholder involvement. To establish a solid ground, data quality needs to be assured early on to limit uncertainties and get an early evaluation of the project's feasibility.

It is also found that setting and managing goals is highly challenging. Measures used to evaluate ML products are often arbitrary and highly relative to the domain's characteristics. In addition, it is often challenging to establish baseline measures to compare against. Goals and targets are also influenced by the high levels of uncertainty in ML projects, which requires significant effort to cope with.

Finally, the study finds that RE for ML is challenged by the technical knowledge and involvement of stakeholders, specifically users of the final product. ML products differ greatly from regular software products, and involving stakeholders early and consistently can both ensure better requirements elicitation, as well as increased understanding and knowledge of ML characteristics.

5.9.2 RQ2: How do the practically observed challenges compare to the theoretical challenges?

The challenges found in the case study are in most cases in line with those suggested in Table 2.1. RE processes are highly impacted and challenged by the strong dependence on data, in combination with uncertainties in many, if not all, parts of development. However, in the exploratory case setting, the challenges suggested by literature regarding legal and ethics were not found as challenging, and there was a lack of data regarding coupling, team collaboration, and requirements relations, possibly due to a lack of RE know-how.

In Table 5.2, the comparison between theoretical challenges and the analytical themes is summarized. As presented in the discussion, some challenges were not found in the study, but further research is required to ensure that the lack of evidence is not due to the specific case context. The theme *Project Purpose* is not mapped to any theoretical challenges, due to the indirect connection to RE.

Table 5.2: Relations between theoretical challenges and themes from the study. Themes in parenthesis represent a weaker relation to the challenge.

Group	**Challenge**	**Themes**
Multi-disciplinarity	Cross-functional Knowledge	Technical Knowledge
	Collaboration Difficulties	(Stakeholder Involvement & Cooperation), (Technical Knowledge)
Field Novelty	Lack of Techniques & Practices	*Not found*
	Ethics & Legal Impact	(Ethics & Legal)
	Req. Types and Relations	(Technical Knowledge)
Uncertainty	Outcome Unpredictability	(Structure & Resources)
	Customer Expectations	Goals & Evaluation, (Structure & Resources)
Coupling	Increased External Coupling	*Not found*
	Increased Internal Coupling	*Not found*
Data Dependence	Evolution & Retraining	Data & Domain
	Data Quality Issues	Data & Domain
	Domain Knowledge Dependence	Data & Domain
	Domain Specific Viability	Data & Domain
Comprehension	Achieving User Acceptance	Stakeholder Involvement & Cooperation, Technical Knowledge
	Evaluation, Acceptance & Metrics	Goals & Evaluation, Technical Knowledge
	Missing Problem Definitions	Technical Knowledge
	Model Interpretation Difficulties	Technical Knowledge

5.9.3 RQ3: What are the characteristics of successful and unsuccessful ML projects and processes?

Concurring with the literature background, data quality is found to be one of the major success factors of ML projects. Low-quality data impacts the product outcome and can also impact the requirement and development process since significant time is spent to improve, interpret, or structure the data.

Further, it is found that a clear problem definition and goal statement is perceived to increase the likelihood of success. It improves the evaluation process, which in turn makes sure that the decision-making process is well-functioning, that successful projects continue, and that less successful projects are discarded or paused. The reason why the projects should be started needs to be defined early on. It could either be to create value for a customer in the short term, for example, by developing a product and implementing a solution, or the reason can be to examine, try, and learn about techniques or new research. The latter can create value in a long-term perspective, and should not be evaluated on the project's short-term business contribution.

5.9.4 RQ4: What commonly used techniques and practices for RE could be suitable for ML projects?

Due to the case organization's low experience in RE practices, suggestions for new or improved practices were not included in the interviews or the focus group. The suggested practices have instead been derived from the challenges and success factors, practices suggested in the literature, and the authors' experience in the field of RE. However, further validation of the positive impact of these practices is needed. In Table 5.3, the suggested practices are summarized and connected to each theme. Due to the lack of observations of ethical and legal challenges, no practices are suggested. RE practices for the theme *Project Purpose* are also lacking, since the theme is related to the organizational visions, rather than project requirements. However, purpose reflections on a project's purpose can greatly impact the RE processes.

Firstly, use case diagrams could be used to elicit requirements and manage customer expectations in a better way. Awareness of the users' relation to the suggested product can be beneficial both for the development team and the user group. Secondly, a scenario-based RE process can strengthen the expectations management, as well as provide information on the user maturity, aiming for a suitable, sufficiently explained, and usable product. Thirdly, the case organization is advised to increase compliance with their Quick Evaluation practices, as it is a simple and suitable tool to predict challenges and tailor actions to the project's unique situation. Lastly, to cope with the challenges of goals and evaluation, as well as uncertainty management, the study suggests the use of ML-GORE to highlight uncertainties and apply an empirical approach to the goals.

Table 5.3: Suggested RE practices for identified themes.

Theme	Practice
Data & Domain	Pyramid Model, Quick Evaluation
Ethics & Legal	N/A
Stakeholder Involvement & Cooperation	Scenario-based elicitation
Structure & Resources	Mode 1 & 2 differentiation
Goals & Evaluation	ML-GORE, Use case diagrams
Technical Knowledge	Scenario-based elicitation, Quick Evaluation
Project Purpose	N/A

5.10 Delimitations and Validity

Following a description of the delimitations of this thesis, this section presents potential threats to the validity of this study. The validity discussion is separated into internal-, external- and construct validity, and also presents mitigating strategies used throughout the thesis.

5.10.1 Delimitations

The study has mainly considered the process after empathy and ideation, and before prototyping and deployment. Hence, the study might be limited in its analysis of challenges regarding monitoring, deployment, ML- or DevOps, and other product-related practices. This could have impacted the team's notion of coupling challenges, which might present themselves when deploying products as integrated parts in larger systems.

The study is delimited to a single type of process within one organization and is not compared or verified with other organizations. Therefore, one can not conclude that the result of this study is valid in its entirety for other organizations. However, it might be valid for similar contexts.

When collecting data through interviews and a focus group, the selection has been delimited by involvement or connection to the process in the study. Therefore, no data or opinions have been collected from sources that are entirely external to the process. This could have strengthened the analysis of general user maturity in relation to ML but was considered out-of-scope when studying the case process.

5.10.2 Internal validity

The purpose of the case study is to describe phenomena in the context of the case. Therefore, the study tries not to draw conclusions regarding the causes of the challenges, but rather to describe and find evidence of the existence of challenges.

Further, the study does not aim to statistically validate the cause-and-effect relationship between a factor and the rate of success, but merely observes and conclude what participants in the study perceive as success or failure factors. Given the nature of case studies, the ability to condition on, or rule out, a single parameter under study is nearly impossible due to the interrelationship between the proposed success factors, the process, and the people involved in the study.

5.10.3 External validity & Generalizability

The statistical generalizability of case studies is inherently limited by the context, and one should be careful not to draw general conclusions based on this study. However, results from this study can contribute to the knowledge about these specific contexts. Differences or similarities with the empirical results could depend on organizational differences or similarities with other case studies performed. Hence, it is questionable whether the RE for ML challenges differ, or if the contexts differ in such a way that the challenges are affected.

As the core of the products developed by the team, as well as frameworks and techniques, are not specific to either the automotive industry or supply chain, the generalizability of the result should not be limited to these domains. The Exploration Mode Development utilized by the team could affect the generalizability of the findings, since this tailored process could differ from other organizations. The key aspect of this practice is the focus on exploring, applying aspects of design thinking to achieve this. It might be more similar to R&D than more typical development practices. However, the elements of Exploration Mode development are largely a composition of existing frameworks, and should not significantly differ from other processes that are based on design thinking and agile workflows. Lastly, it is important to consider the case description and process description before comparing or using the results of this study in other contexts.

5.10.4 Construct validity

The empirical data on challenges from interviews can contain threats to construct validity, as the challenges are described as perceived by the interviewees. Different roles can have vastly different challenges, and employees in a small team might also be reluctant to disclose challenges that might impact their employment. This is counteracted by the use of additional data sources, such as observations and reading of documentation, as well as a combined focus group and questionnaire.

The use of a focus group, combined with a questionnaire, to confirm the results is positively affecting the construct validity. The answers from an interviewee could be influenced by several factors, for example, a previous meeting that day or nuances in how the questions were asked. Using a focus group, potential misinterpretations or misunderstandings can be corrected, and raising statements in a group setting provides data on the general understanding and perception of success factors and

challenges. This notion of "aided recall" is supported by Höst et al. [28], discussing the benefits of focus groups, as participants can collectively discuss and remember aspects of the questions and answers. It also helps to identify nuances or edge cases that would skew the results. The presence of managers in the focus group might influence the other participants' responses, which might be inclined toward the company line or ambition. However, the general perception during the study has been that the management is open to constructive criticism, and tries not to limit any opinions toward current processes.

While the focus group and questionnaire were done in the same session, they were kept separate to decrease the influence of the discussion on the questionnaire responses. First, the entire questionnaire was answered, and then the anonymous results were used as discussion material in the focus group, to avoid participants adapting to the perceived consensus opinion when answering the questionnaire.

The team in this study has generally low knowledge of RE principles and practices. This could influence the data collection since questions had to approach the topic with minimal use of RE-specific terminology. To mitigate the impact of this, data collection was mainly done in face-to-face settings where terminology could be explained and questions could be raised by the interviewee or participant.

The pairwise coding of the interview transcripts assures quality and agreement in the coding process. However, an improvement of the study could have been to measure the agreement between coders. Upon disagreement when coding a text snippet, the different aspects were discussed and a mutually agreed code was chosen. This was not documented, which would have been beneficial to discuss the aforementioned agreement during coding.

The case organization has been discussing its processes and practices for the duration of the thesis. The purpose of the thesis is to contribute to this discussion, but some thoughts and reflections have already started to appear as suggested changes to the processes. This could influence the data collection since the questions regarding past projects might be answered from the perspective of an already-changed organization. However, the suggested changes had not been implemented at the time of the interviews, which suggests that identified challenges and success factors are still valid. Further, the description of the case depicts the process as it was during the data collection, not how it is planned to be in the future.

A further threat to the construct validity could be the potential bias in interview questions. It is possible that unnoticed skews in objectivity might have influenced the way interviewees interpreted and answered questions. To mitigate this effect, the interview questions were peer-reviewed by the supervisors from Chalmers and the industry. A further threat to validity concerning the interviews is that the theoretical framework was established before the interviews. This might have induced confirmation bias when coding and examining similarities between literature and responses. Lastly, the interpretation and analysis of the interviews could have been

influenced by the use of two languages, and translation (Swedish to English), where subliminal meaning can be lost. The use of multiple languages was however motivated by the strength of, when possible, using the interviewees' first language. The interviewers' language abilities were also considered high enough to mitigate risks in translation.

5.11 Further research

This study identifies challenges in RE for ML and suggests different practices that might mitigate their impact. However, studies to validate these practices, and empirically support their effect, are needed.

Further, it is suggested to investigate other phases of stand-alone ML development. This study finds differences between stand-alone deliveries and integrated systems (which are more present in previous research), but the scope was confined to the early, explorative, phases of development. There might be further insights if the deployment, and long-term maintenance, perspective is considered as well.

Further studies could also benefit from comparing organizations of different SE and RE maturity. In this case, the RE know-how of the organization was relatively low, and the team composition was weighted towards data science and statistics, rather than software engineering. This might have induced challenges, or abilities, that would be different in a SE-specialized organization.

6
Conclusion

ML components are often part of larger software systems, and organizations need to adapt their requirements processes to the new challenges that come with RE for ML. However, there is a lack of empirical data in the emerging field of ML research, specifically RE for ML, and it is relevant to extend the chain of evidence to find new best practices specifically incorporating RE for ML. The purpose of this thesis was to provide empirical evidence on challenges with RE for ML, and to contribute to the general knowledge within this field.

In this thesis, a case study was performed at Volvo Group, collecting data through interviews, observations, and documentation extraction. This data was evaluated through a focus group and questionnaire for further data validation. Results show that data quality is central to ML project success, and suggests that data requirements are formulated early, and separately, to assure quality to some extent before committing to a project. Further, both the involvement and inherent competencies of the stakeholders are shown as contributors to success.

This study concurs with theoretical challenges in most cases, and provides further empirical data to support them. However, the study did not present enough data to support theoretical challenges with ethics, legal, and coupling aspects. Challenges regarding the purpose of projects are also presented in addition to the theoretical ones.

Lastly, the thesis suggests practices for further validation, that could mitigate challenges and increase chances of success. These practices include the Pyramid Model, scenario-based requirements elicitation, ML-GORE, and use case diagrams. Recommendations specific to the case organization include increasing compliance on Quick Evaluations and ensuring a thorough decision process between Mode 1 and Mode 2 development.

Bibliography

[1] What is safe - framework for business agility, Nov 2022.

[2] Christof Ebert and Maria Paasivaara. Scaling agile. *IEEE Software*, 34(6):98–103, 2017.

[3] Elizamary Nascimento, Anh Nguyen-Duc, Ingrid Sundbø, and Tayana Conte. Software engineering for artificial intelligence and machine learning software: A systematic literature review. *CoRR*, abs/2011.03751, 2020.

[4] Fuyuki Ishikawa and Yutaka Matsuno. Evidence-driven requirements engineering for uncertainty of machine learning-based systems. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*, pages 346–351, 2020.

[5] Fuyuki Ishikawa and Nobukazu Yoshioka. How do engineers perceive difficulties in engineering of machine-learning systems? - questionnaire survey. In *2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP)*, pages 2–9, 2019.

[6] Hugo Villamizar, Tatiana Escovedo, and Marcos Kalinowski. Requirements engineering for machine learning: A systematic mapping study. In *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 29–36, 2021.

[7] Ralph Rowland Young. *The Requirements Engineering Handbook.* Artech House Technology Management and Professional Development Library. Artech House, Inc, 2004.

[8] Mario Sergio Salerno, Leonardo Augusto de Vasconcelos Gomes, Débora Oliveira da Silva, Raoni Barros Bagno, and Simone Lara Teixeira Uchôa Freitas. Innovation processes: Which process for which project? *Technovation*, 35:59–70, 2015.

[9] Keith Pavitt. 86 Innovation Processes. In *The Oxford Handbook of Innovation.* Oxford University Press, 01 2006.

[10] Gartner Inc. Definition of bimodal - gartner information technology glossary.

[11] Tim Brown. Design thinking. *Harvard Business Review*, 86(6):84 – 92, 2008.

[12] Ashley Aitken and Vishnu Ilango. A comparative analysis of traditional software engineering and agile software development. In *2013 46th Hawaii International Conference on System Sciences*, pages 4751–4760, 2013.

[13] Paul Flewelling. *The the Agile Developer's Handbook : Get More Value from Your Software Development: Get the Best Out of the Agile Methodology.* Packt Publishing, Limited, 2018.

[14] Soren Lauesen. *Software requirements : styles and techniques.* Addison-Wesley, 2002.

[15] Irum Inayat, Siti Salwah Salim, Sabrina Marczak, Maya Daneva, and Shahaboddin Shamshirband. A systematic literature review on agile requirements engineering practices and challenges. *Computers in Human Behavior*, 51:915–929, 2015. Computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era.

[16] Lan Cao and Balasubramaniam Ramesh. Agile requirements engineering practices: An empirical study. *IEEE Software*, 25(1):60–67, 2008.

[17] Rashidah Kasauli, Eric Knauss, Jennifer Horkoff, Grischa Liebel, and Francisco Gomes de Oliveira Neto. Requirements engineering challenges and practices in large-scale agile system development. *Journal of Systems and Software*, 172:110851, 2021.

[18] Görkem Giray. A software engineering perspective on engineering machine learning systems: State of the art and challenges. *Journal of Systems and Software*, 180:111031, 2021.

[19] Anders Arpteg, Björn Brinne, Luka Crnkovic-Friis, and Jan Bosch. Software engineering challenges of deep learning. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 50–59, 2018.

[20] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300, 2019.

[21] Zhiyuan Wan, Xin Xia, David Lo, and Gail C Murphy. How does machine learning change software development practices? *IEEE Transactions on Software Engineering*, 47(9):1857–1871, 2019.

[22] ISO25010. ISO/IEC 25010 - System and software quality models. Standard, International Organization for Standardization, Geneva, CH, 2022.

[23] Hrvoje Belani, Marin Vukovic, and Željka Car. Requirements engineering challenges in building ai-based complex systems. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 252–255, 2019.

[24] Hans-Martin Heyn, Eric Knauss, Amna Pir Muhammad, Olof Eriksson, Jennifer Linder, Padmini Subbiah, Shameer Kumar Pradhan, and Sagar Tungal. Requirement engineering challenges for ai-intense systems development. In *2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)*, pages 89–96, 2021.

[25] Andreas Vogelsang and Markus Borg. Requirements engineering for machine learning: Perspectives from data scientists. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, pages 245–251, 2019.

[26] Jennifer Horkoff. Non-functional requirements for machine learning: Challenges and new directions. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 386–391, 2019.

[27] Khan Mohammad Habibullah, Gregory Gay, and Jennifer Horkoff. Non-functional requirements for machine learning: understanding current use and challenges among practitioners. *Requirements Engineering*, 28(2):283–316, 2023.

[28] Martin Höst, Austen Rainer, Per Runeson, Bjorn Regnell, and Björn Regnell. *Case Study Research in Software Engineering : Guidelines and Examples.* John Wiley & Sons, Incorporated, 2012.

[29] Lawrence A Palinkas, Sarah M Horwitz, Carla A Green, Jennifer P Wisdom, Naihua Duan, and Kimberly Hoagwood. Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5):533–544, 2015.

[30] Johnny Saldaña. *The Coding Manual for Qualitative Researchers.* SAGE Publications, Ltd, 2013.

[31] Eleonora Giunchiglia, Fergus Imrie, Mihaela van der Schaar, and Thomas Lukasiewicz. Machine learning with requirements: a manifesto, 2023.

[32] Douglas Cirqueira, Dietmar Nedbal, Markus Helfert, and Marija Bezbradica. Scenario-based requirements elicitation for user-centric explainable ai: A case in fraud detection. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 321–341. Springer, 2020.

A

Interview Templates

Interview Template Round 1 Version 1

Metadata

General information

- Date of interview
- Location
- Interviewers
- Language

Interviewee information

- Agree to audio recording?
- Name
- Role
- Time of employment
- Closest manager
- Field of work
- Years in field of work

Description of thesis project The purpose of the study is to provide empirical evidence to contribute to the existing body of knowledge surrounding general Requirements Engineering (RE) and RE for Machine Learning (ML). The purpose is further to identify relevant RE practices, and either corroborate or discard the suitability of these in the field of ML development.

The first round of interviews will serve as an initial data collection on current practices and procedures, as well as success- and failure factors, challenges, and solutions.

There might be an additional round of interviews or workshops depending on the outcome of the first round.

Case context

- What is your usual role in an ML related exploration/project?
- Does your role or responsibilities change in different ML explorations?
  - In what way?
- How are you involved in setting goals, targets or measures to determine an exploration's success?
  - Can you describe the process from your point of view?

- In what way do you evaluate project success?
  - How to determine if these goals, targets or measures have been met?

Ongoing projects

- If relevant, which ML explorations are you involved in right now?
- What role do requirements play for you in those projects?
  - Where do they come from? A specific role/person/group?
- In what way are you working with, or act as, a stakeholder in that/those explorations?

Past projects

- In your opinion, what are significant characteristics of a successful ML project?
- In your opinion, what are significant characteristics of a failed ML project, or learning?
- In your opinion, how much do ML projects differ from each orther?
  - Organization
  - Documentation
  - Requirements
  - Exploration
- Do any of your past ML projects stand out to you as exceptionally successful or unsuccessful?
  - Evaluate the causes. Challenges or successes?

Challenges in Requirements Engineering

- Are there any key challenges related to requirements, goals, targets or measures that have been present in current or past projects?

Interview Reflection

The first two questions were only asked to the first two interviewees of round 1.

- Did you understand the questions and terms used?
- Did you find the questions or introduction leading?
- Do you have any further reflections or input to the case study?

B

Observation Template

Table B.1: Observation template with examples.

Notation	**Example**
Observation ID	O1
Date & Time	2023-01-25 15.47.57
Context	Decision meeting
Actors	AdA Team
Observation	KPI:s with targets but large fluctions in actual results. Are the targets really relevant?
Sentiment	Negative
Conclusion/Hypothesis	
References	Link to notes

C
Code Book

Table C.1: Complete code book with both a-priori and emerging codes.

Code	Group	Type	RQ
UACC	Comprehension	Challenge	RQ1, RQ2
User acceptance and usage when comprehension of ML is low			
UMATU	Comprehension	Challenge	RQ1, RQ2
Challenges with maturity levels and knowledge in the user/customer group, prohitibing adoption and usage.			
PROBDEF	Comprehension	Challenge	RQ1, RQ2
Problem definition. Crucial to comprehend the problem to make a good model			
MODINT	Comprehension	Challenge	RQ1, RQ2
Interpretability is a challenge relating both to comprehension, but also NFRs and ethics.			
INNOV		Challenge	
Challenges connected to finding innovations and disruptions.			
EV&ACC&MET	Comprehension	Challenge	RQ1, RQ2
Evalutation, acceptance and metrics decisions			
EXTCOUP	Coupling	Challenge	RQ1, RQ2
External coupling - reliance on other systems.			
INTCOUP	Coupling	Challenge	RQ1, RQ2
Internal coupling - changing something changes everything			
EVO&RE	Data Dependence	Challenge	RQ1, RQ2
Evolution & Retraining. Accounting for maintenance processes in RE			
DATQUAL	Data Dependence	Challenge	RQ1, RQ2
Data quality requirements challenges			
DOMKNW	Data Dependence	Challenge	RQ1, RQ2
Domain knowledge is critical to understand data features			
DOMSPEC	Data Dependence	Challenge	RQ1, RQ2
Domain specification might have a larger impact on ML RE, since model performance is context dependent			
CROSS	Multidisciplinarity	Challenge	RQ1, RQ2

Continuation of Table C.1			
Code	**Group**	**Type**	**RQ**
Cross-functional knowledge as a challenge in ML-dev. and RE			
COLL	Multidisciplinarity	Challenge	RQ1, RQ2
Collaboration with stakeholders.			
TECH&PRAC	Field Novelty	Challenge	RQ1, RQ2
Lack of best practices techniques in the field			
ETHIC	Field Novelty	Challenge	RQ1, RQ2
Ethical challenges arising in the field of ML SE and RE			
REQT&R	Field Novelty	Challenge	RQ1, RQ2
Requirements types and relations			
OUTPRED	Uncertainty	Challenge	RQ1, RQ2
Outcome predictability is lower in ML, how to conduct RE and specify outcome correctly?			
CUSTEXP	Uncertainty	Challenge	RQ1, RQ2
Managing customer expectation in regards to RE, while understanding uncertainty			
GOALSPEC		Success/Failure	RQ3
DATA		Success/Failure	RQ3
INVOLVE		Success/Failure	RQ3
BASELINE		Success/Failure	RQ3
REASON		Success/Failure	RQ3
STORY		Success/Failure	RQ3
EXPECT		Success/Failure	RQ3
OWN/SPONS		Success/Failure	RQ3
HANDRED		Success/Failure	RQ3
TIME		Success/Failure	RQ3
KNOWHOW		Success/Failure	RQ3
GOALS&EVAL		Practice	RQ3, RQ4
Descriptions of practice when setting goals or targets.			

Continuation of Table C.1

Code	Group	Type	RQ
DOC		Practice	RQ3, RQ4
Observation or comment regarding documentation procedures or practices.			
CULT		Practice	RQ3, RQ4
Mentions of cultural aspects of the case, team, people or processes			
DM		Practice	RQ3, RQ4
Decision making. Comments or observations of decision making actions and/or processes.			
STRUC		Practice	RQ3, RQ4
Structure. Comments or observations regadring the organizational structure of the case context.			
EXPL		Practice	RQ3, RQ4
Exploraiton. Comments or observation of practices regarding explorations or creative development, brainstorming, etc.			
IMPRO		Practice	RQ3, RQ4
Suggestions on improvments in current practises.			
COOP		Practice	RQ3, RQ4
Cooperation and/or collaboration with stakeholders.			
EXAMP		Practice	RQ3, RQ4
ROLE		Metadata	
Role descriptions and comments regarding field of work etc.			
EXPE		Metadata	
Experience. Comments on teams or induviduals experiences and previous knowledge			

D

Interview Code Condensations

As explained in subsection 3.5.1, the interviews have been summarized for each code.

D.1 Practices

GOALS&EVAL - Goals, Targets and Evaluation
Some interviewees said that explorations start with a discussion around the value statement and what goals to reach. However, other answers say that vague goals appear during the exploration cycle. There are suggestions that the number of a priori goals or requirements depends on project size. One interviewee is also adamant that there should be different levels of requirements depending on the project status (empathize, define, ideate, etc.), and align the expectations on each status segment.

Sometimes goals are added to the DevOps board in Azure, but sometimes it is just "forgotten".

It's not uncommon to formulate goals that just state "improvement compared to current" in terms of some business parameter. Statistical or model-related goals are rarely presented to the stakeholders unless there is no business metric to present. They are, however, measured by data scientists but can mean vastly different things (e.g. accuracy). A reasonable metric could be how fast the solution can be implemented, according to one interviewee.

A domain expert mentioned that when measuring for example forecast improvement, the measure has to be carefully selected based on the domain values. Even though goals are mostly set as improvements to the current state, they could be measured in many different ways.

Most interviewees stated that in the end, business values should drive both goals and evaluation. Goals can come from higher up in the organization, from the team, or from stakeholders. When goals or requirements are set within the team, the different roles can provide different expertise. It's good if one can take the domain/business side and one can take the technical side. It might be hard for stakeholders to set goals especially when they are new to analytics, then it can be done in cooperation with AdA. AdA presents suggestions. Goals and requirements might develop as they go and learn more about the problem or related processes, or when specific stakeholders get involved.

When it comes to evaluation, interviewees mention that it's much easier to evaluate specific and defined goals, but that they are harder to formulate. They are sometimes bad at setting goals, and instead test and evaluate as it goes.

A standardized quick evaluation should, according to the process, take place when defining an exploration. This can help identify which challenges need to be addressed before developing. There are also value estimations, but one interviewee mentioned that they are bad at following up on the actual value created by the explorations.

There is one example of thorough evaluation through simulation, where the model could be evaluated on prediction and sensitivity etc.

A bit separated from the product success evaluation, there are mentions of learning success in the sense that the main goal of an exploration can be a "learning experience". This can be relevant if the receiver organization is immature in terms of data and/or analytics. Management interviewees clearly focus more on this parameter, connecting to organizational transformation. It could also be valuable to set goals for how to build competencies out of a scrapped project. This could provide value in future explorations, according to one interviewee.

DOC - Documentation Practices
The teams have a formal structure for documentation with cards, demo sessions, powerpoints, and a confluence page. However, it is mentioned that the content in the documentation might vary depending on the writer. Depending on the role they might focus on different things. It is mentioned though that exploration leads have added a level of structure in way of work to all projects that do not remove too much freedom.

It is also mentioned that high-quality documentation is more likely when the goal is to develop a specific product, compared to more exploratory work. Also, if it only ends up with a learning, documentation might be of less quality or less extensive.

One domain expert does not believe ball-park value adds value for him/herself but might do for others.

No specified owner of documents/artifacts, and no one is responsible for them being updated and revised. Can still find information about previous projects that would be good to know, that have fallen between the chairs. Might then redo work already done. Bad at evaluating "way of working" on all levels in a structured manner.

They have guidelines on documentation and code but do not do quality control.

CULT - Cultural Aspects
It is mentioned that there is a culture of request delivery. AdA is seen as a support function that you can hand over a problem to and then get a solution, customers

do not always realize that they are expected to participate. There is also a culture of that they do not need to handle or maintain any automated models, that is the role of the IT department. This could be due to a general lack of knowledge of data and analytics.

Focus is on creating a culture of explorative thinking. Largely dependent on leadership, requires a more free less hierarchical leadership but that does not exist in all parts of the organization. It is also mentioned that to facilitate this focus can not only be on value creation but instead on learning. It is mentioned that they are not always good at following standards, but instead, they are more creative. The creative process is spreading in the organization and teams outside AdA have adopted it.

It is also mentioned that AdA might sometimes be too quick to accept projects, without investigating if it is actually possible. Concerning projects, it is mentioned that they have a hard time closing started projects and that for a project to move forward and be successful it requires sponsorship or support from the organization.

Might be friction when moving on to implementation. It is then needed to have more strict requirements on for example code quality, and that might be lacking at that stage.

Might be a culture of implementing "way of working" for the wrong reasons, not because it is actually needed but instead because it is “in” or“cool”.

Data Scientist gets programming questions. People do not dare to voice their opinions.

DM - Decision Making

Currently, the formal exploration process contains multiple decision points on whether to continue or not. However, the conditions and ways of evaluation are unclear. The development team often has to define success themselves, and the success feasibility evaluation is subjective.

There are examples of explorations that have been going on for years, even though there is no perceived progress. Interviewees note that there are problems regarding closing explorations and using tangible requirements for decision-making. For example, if an exploration does not pass an initial data accessibility evaluation, it should not continue.

Official process: The process owner is responsible for a process and coordinates a forum where you gather all stakeholders and there take decisions on if explorations have met the goals. Can get a sign-off from stakeholders that it is ready to use. In practice: Some people do that others do not. Mostly the most engaged person that takes the decision.

Also if they see improvements they can start using it directly without the formal process. Then they share that they have done it.

STRUC - Project and Team Structure
Explorations differ quite a lot from each other. Some interviewees even state that most explorations are unique. There is a formal structure that is the same for all explorations, but in practice, it differs.

- Preparation: Shorter exploration (e.g. pulse lab) requires more preparation to achieve value in that short pulse.

- Staffing: The teams are created on an initiative basis, and teams rarely look the same.

- Time: Some explorations are very short, and some have a large scope and consist of sub-explorations.

Mostly, the team consists of (at least), one data scientist, one SCAE, and one exploration lead. Sometimes a stakeholder as well. There are mentions that explorations often lack ownership, and that things lose structure since prioritization is mostly interest-based.

Regarding different modes of development, interviewees find that the team is not very good at differentiating between Mode 1 and Mode 2 (exploration). Usually, all ideas are implemented through mode 2.

EXPL - Exploration, Innovation, and Creativity

The team basically has two types of explorations, one where the idea or problem is clear and one where they just find a technique or something interesting and tries it out. Generally, with the first one a stakeholder comes with a problem and the second on the teams tries to find a use case. Both are to some extent explorative, but the latter on a bit more. Can be a creative process together with a stakeholder in formulating the problem or requirements.

The level of exploration is higher in the early stages of a project, and the requirements are fewer or less specific. Increases later but it is not a clear process of how and when. Too much structure both in terms of requirements and processes might hinder creativity and can then miss important aspects or values. The goal of explorations should take 30 days might hinder the creative process.

Needs to dare to try new techniques and methods and give them time.

Explorations should sometimes be less explorative and treated more as regular projects (mode 1).

They learn something new in every project, a large degree of exploration that might

mean that it goes slower, but it builds competencies.

COOP - Cooperation with Stakeholders
The technical know-how of the stakeholders usually determines who's leading or driving an exploration, as well as their overall involvement. In some cases, stakeholders are involved in the development, and in some cases, they aren't. In the latter case, it's more of a request-follow-up-delivery process. In most cases, however, there are discussions around defining the current state and the problem.

Interviewees also mentioned geographic and organizational location as involvement factors. It can also depend on the project/problem definition, and whether stakeholder involvement would enhance value.

Request-delivery processes might decrease the possibility for the user/customer to receive and use the solution, due to poor technical know-how and low understanding of the product. Though the project managers try to inform about the issue, interviewees find that projects rarely stop due to low involvement. Instead, they are deployed by AdA in order to create value.

AdA is no IT organization when it comes to maintaining and monitoring solutions. That competence is needed in the receiver organization after delivery.

D.2 Challenges

COLL - Collaboration
Collaboration with stakeholders is difficult but necessary for success. They sit on the domain knowledge and the demands should come from them and they should be the ones evaluating if it is good enough. The difficulty lies in time and resource constraints with the stakeholders and their technical and data expertise. Administrative difficulties can also appear when many are involved in the development. There is therefore difficulty in determining when to involve the stakeholders and to what degree.
There are also cultural aspects to the challenge of collaboration. There exists sometimes a mindset that AdA should just solve the problem for the customer, and not together with them. This in one case mentioned as a result of not focusing on the most pressing problem for the stakeholder. This can also affect the interest, which makes the handover more difficult.

UACC - User Acceptance
Mentioned by two interviewees:
Customers are often used to doing things a certain way, and even if it sometimes has problems, it works decently. And it is a process that they are responsible for. They then need to trust the result of the model, and it could be a challenge to get them to do that. This could be done by explaining how the model works.

UMATU - User Maturity
Several interviewees have mentioned that there is a challenge with customers/stakeholders not having the technical knowledge. When there is time to hand over the product it becomes difficult when the customer does not know how to run it or maintain it. There is a bridge that needs to be crossed, either by the developers or the customers. A balance between making the product easy to use and raising the technical knowledge of the user.
It is also a challenge earlier in the process. If the stakeholders are not used to thinking of data, it is hard for them to formulate the problem and define the goals. It is also more difficult for them to interpret the results.
There is also a tradeoff between learning and developing, as time spent on either takes time away from the other. The development becomes less effective with more people in the team.

PROBDEF - Problem Definition
Not having a defined idea can be challenging. Interviewees state that the specified metric goal of the model implementation is one part, but another important challenge to solve is where to go afterward and what the actual problem is that they are going to solve. A project can also be negatively affected by not knowing who the recipient is, and what they are going to do/solve with the product.

INNOV - Innovation
Finding engagement for disruptive innovation is hard. Solving existing problems creates more involvement, rather than trying new techniques, which can create distance.
But if you are to find disruptors, one interviewee argued that you cannot just work on what you already know.

EV&ACC&MET - Evaluation, Acceptance & Metrics
Interviewees mention that an absence of goals or targets makes it harder to determine success and evaluate results. Spoken or undefined targets can also result in adjustments of goals to adapt to what has been achieved.

Interviewees have also found it challenging to establish a baseline for current practices. This, in turn, makes the evaluation more challenging since "good" forecasting or prediction accuracy can be arbitrary, or at least relative to the current state. One interviewee also mentioned that statistical evaluation is lacking and that there is a need to account for pure luck. Another interviewee also mentioned that you have to avoid "cheating" when evaluating, for example looking at future values.

They also find it challenging to translate improvements into business value and to select which measures to look at when evaluating performance. Especially if the organization or current process is not "digitally mature". Evaluating business value is even more challenging if you just want to explore something unknown.

Exploring teams might also be too eager to start, forgetting to scope and define the

problem. It is also challenging to know when to stop, either to kill an exploration or move to a prototype in order to create value. Killed projects are also lacking in terms of post-evaluations, which could work as a future reference.

Lastly, they mentioned that there is a challenging tradeoff between specific goals and creativity. Maybe there should be a proportional relationship between requirements and scope, and increase requirement specifications as the project progresses.

INTCOUP - Internal Coupling
One interviewee mentioned this: The code is rarely modular, which hinders reuse. Could depend on the coding skills of data scientists.

EVO&RE - Evolution & Retraining
One interviewee mentioned this: Retraining and modification due to changing external parameters might be a challenge. The potential need for scenario planning, when to modify models.

DATQUAL - Data Quality
Interviewees mention that data is central and that this poses several problems. First, it is a question of if the data even exists in the first place. Secondly, it has to be made available at the beginning of the project otherwise a lot of time is spent on acquiring it. This is a problem within Volvo as it is not always easy to get access to the data. It is also mentioned that even when you get access to it it might only be a snapshot of the data, and it is therefore difficult to test the viability of if it turns into a product. Thirdly, it is that the quality of the data has to be ensured.

There is also a problem with understanding the data and the context. There exist situations where the data comes from unstructured processes that differ between parts of the organization, which limits the viability of the final solution.

Sometimes they still try to develop even when they do not have enough data.

DOMKNW - Domain Knowledge
The domain knowledge is approached by interviewees in two ways, data and business/value.

Knowledge about the current state of the operations is crucial to evaluate if a product is better or worse, compared to a baseline. The same goes for evaluating what good performance means in terms of accuracy and statistical measures. Lastly, the improvements need to be put into context. What business value does a 2% improvement create? It's a challenge of translation.

Domain knowledge is also important for data understanding. Where, how, and when is the data measured, is it measured the same everywhere, and does the same data points mean the same in all contexts?

DOMSPEC - Domain Specification
One interviewee mentioned this: Different users of the same product/model might operate in different contexts. It is challenging to make sure that the model is good for the required context, time frame, etc.

CROSS - Cross-Functional Knowledge
Interviewees find that distributed knowledge is important but challenging. One expert cannot do everything, and the community needs the skill to operate products. Further, resource prioritization decisions are a challenge when collaborating outside the team.

This becomes apparent for Data Scientists, who often work on software engineering or developer tasks. Largely due to that SEs are prioritizing larger projects or systems.

TECH&PRAC - Techniques & Practices
Two interviewees mentioned that current practices (for example agile), or large frameworks, are not suitable for explorative development which is closer to R&D. Complicated frameworks limit the understanding, and perhaps ML development cannot be a streamlined process.

A third mentioned the risks with frontloading requirements and suggested incremental requirements process throughout the exploration/project.

ETHIC - Ethics & Legal
One interviewee mentioned this: For certain areas, in this case customs, there are limitations to the decision-making possibilities of the models since there could be legal implications.

OUTPRED - Outcome Unpredictability
Interviewees mention that there are several challenges with uncertainty. The success of the outcome is difficult to estimate beforehand and therefore causes challenges in setting requirements and goals and estimating the timeframe. This is especially relevant when utilizing new technologies and with unclear problems. One case is mentioned where the model developed was not good on its own, but together with the current method it worked better. The goals could then be not too specific because then you might have discarded the project as a failure.

It is also a problem when estimating the potential value of the project. But then it is more a question of estimating to what extent it can be deployed, whether it will be one site or the whole world. But also when the value it creates is not tangible like man hours.

It was also mentioned that there is a problem with people not understanding that ML is more similar to R&D than other types of development.

CUSTEXP - Customer Expectations
Interviewees mention that managing customer expectations is crucial but challenging. The first challenge is to get the right expectations on what actually can be done both in this specific project but also with ML in general. 100% accuracy can never be reached. The second thing is in what time frame, if it is a new technology it might be difficult to estimate and that needs to be made clear. The third thing is that the customers should know what will be demanded of them in the delivery and application of the model.

D.3 Success factors and stories

BASELINE - Establishing Baseline
Interviewees find that in order to know what to reach, and to properly find the ways to reach it, a baseline comparison value has to be established. This is usually formulated as "better than the current method", but there is a need to relate it to one or several measures with a "status quo" value.

GOALSPEC - Goal Specification
It is mentioned by several interviewees the importance of having clear goals before starting an exploration, in other words, what is the general purpose of the exploration and what should be reached. The goals should also be not too broad. The consequences of not having any goals are that it is difficult to know when you are done or when you have reached a result that is acceptable. It can lead to the exploration dragging on for a long time or the focus on value creation is lost. It is mentioned that it is especially important if the exploration has a specific application area.

INVOLVE - Stakeholder Involvement
The customer or the stakeholder should preferably be involved early on and then continuously in the development process. They should also have an interest in the outcome of the exploration. This has several benefits. The stakeholders sit on the domain knowledge and know what creates value. This could steer the development in the right direction and help to solve potential roadblocks.

If they are involved in the process, it also facilitates the "delivery" of the product. They then know what is expected of them in terms of maintenance and are more caught up on the technical aspects. This is also easier if they are actually interested in the end result, they are then more inclined to spend the necessary resources. However, this all requires time and resources from parts of the organization where that could be limited.

DATA
In order for a project to be successful data must exist, be available, and be of high quality. If it is not available and prepared at the beginning of the project, time is spent on acquiring it which increases lead time. It is also more difficult to estimate

the potential of an idea if you do not know if the relevant data exist or if it can be made available.

It is also mentioned that you need to understand the data, which often requires knowledge about the domain it comes from, otherwise you do not know what to do with it or it can be used incorrectly and give misleading results.

REASON - Reason & Purpose
Interviewees mention that there needs to be a valid motivation or reason behind an exploration. It is mentioned that there is always important to have a business context and not do things just because a new cool technique or solution has been found, and now we need to apply it to a problem. In contrast, however, is also mentioned the importance of learning, and that it can be limiting to only look at the business value and not any other aspects that can be gained from a project.

EXPECT - Customer Expectations
Interviewees find that when working with a customer/stakeholder their expectations on the projects need to be properly managed. Specifically, they mention the time to delivery, as in projects involving new technologies it is difficult to correctly estimate a time frame and this needs to be clear to the customer/stakeholders.

OWN/SPONS - Owner- & Sponsorship
Comments regarding owner- and sponsorship have two perspectives.
Some interviewees mentioned the need for someone to steer and keep the project on the path, and act as a motor toward the end goal.
The other perspective argues for a sponsor allowing explorations and giving the organization time to learn, as well as resources to do so. Sponsors and leaders are also important in establishing an explorative and less business-result focused culture. One interviewee also mentioned that the middle-management section has the longest way to go regarding insights and culture.

HANDRED - Handover Readiness
Users/Customers must be ready for deployment and handover of the product. This includes monitoring and data streaming. One interviewee commented that this can be eased by structuring user operations properly before deploying a new product.
Other comments suggested that it's a question of organizational maturity, where the user/customer brings ideas and is effectively ready to support the development and own a solution.

TIME - Time Estimation
The lead time of an exploration is important to keep the stakeholders interested, if the development is too slow, they might lose interest.

KNOWHOW - Know-How
Interviewees mentioned that technical competence within the team and organization is necessary. It is often not a problem within the team but possibly with

customers/stakeholders. The technical competence of the team is especially a contributing factor to explorations regarding new technologies. In explorations in general it is important that the right people are involved from the start and that there is continuity in who is involved. It was also mentioned that for several projects that could be regarded as successful, it has been that one person has been able to drive the whole project, which has been possible because that person has had a broad area of knowledge, both technical and domain.

E

Questionnaire

Statements were answered on a 4-step Likert scale from "Strongly Disagree" to "Strongly agree".

Table E.1: Questionnaire statements.

ID	Statement
1	**Data & Domain**
1.1	Data quality must be assured before starting an exploration.
1.2	Achieving data quality is a major challenge in explorations.
1.3	Data is unusable without domain knowledge.
1.4	Changing or inconsistent domains is a challenge in explorations.
	Note: Data quality is defined as data of sufficient amount and of high quality and consistency.
2	**Ethics & Legal**
2.1	Models or products have major legal or ethical implications.
2.2	It is challenging to manage legal and ethical aspects.
3	**Stakeholder Involvement & Cooperation**
3.1	Stakeholder involvement is necessary to achieve exploration success.
3.2	Involving and cooperating with stakeholders is challenging.
3.3	Problem-based explorations generate more involvement than technological explorations.
	Note: A "problem-based exploration" is defined as an exploration that is based on a tangible problem that a potential user or customer is facing. The idea might even be initiated by the user/customer.
4	**Structure & Resources**
4.1	A defined owner or sponsor is necessary to achieve exploration success.
4.2	Implementing the model in parallel to defining the data and targets generates more challenges with data quality or problem definition/scope.
4.3	Involving non-technical actors in an exploration decreases development efficiency.
5	**Goals**
5.1	A clear problem statement is necessary to achieve exploration success.
5.2	It is challenging to define specific goals early in an exploration.
5.3	It is challenging to establish a baseline or current state to compare against.
5.4	Requirements and specific goals hinder creativity in explorations.

Continuation of Table E.1	
ID	**Statement**
6	**Evaluation**
6.1	It is challenging to evaluate exploration success.
6.2	Evaluation is improved by having specific goal- or target parameters.
6.3	Explorations should be evaluated mainly by stakeholders.
6.4	Domain knowledge is necessary to evaluate models.
7	**Technical Knowledge**
7.1	Exploration teams have a sufficient competence mix.
7.2	Technical knowledge of the stakeholder is necessary in order to collaborate.
7.3	Technical knowledge of the stakeholder is necessary in order to use the product.
8	**Project Purpose**
8.1	There should always be a specified connection to business value in explorations.
8.2	Technological explorations should not be evaluated on their business value impact.