CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Requirements Grounded MLOps - A Design Science Study

Master's thesis in Computer science and Engineering
Milos Bastajic, Jonatan Boman Karinen

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

Master's thesis 2023

Requirements Grounded MLOps - A Design Science Study

Milos Bastajic, Jonatan Boman Karinen

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

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

Requirements Grounded MLOps - A Design Science Study
Milos Bastajic, Jonatan Boman Karinen

Supervisor: Jennifer Horkoff, Department of Computer Science and Engineering
Examiner: Hans-Martin Heyn, Department of 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

Typeset in LaTeX
Gothenburg, Sweden 2023

Abstract

The use of Machine learning (ML) has increased significantly in recent years, however, organizations still struggle with operationalizing ML. In this thesis, we explore the intersection between machine learning operations (MLOps) and Requirements engineering (RE) by investigating the current best practices, challenges, and potential solutions associated with developing an MLOps process. The goal of this thesis was to create an artifact that would guide MLOps implementation from an RE perspective, resulting in a more systematic approach to managing ML models in production by identifying and documenting the goals and objectives. The study adopted a Design Science Research methodology, which comprised investigating three research questions while the design artifact was being created in parallel. The research questions examined the difficulties currently faced in creating an MLOps process, identified potential solutions to these difficulties, and assessed the effectiveness of these solutions. The study was conducted in three cycles, with each cycle answering all research questions but focusing mainly on one specific question, allowing for the initial creation and subsequent refinement of the artifact based on data collected during each cycle. By establishing a more thorough understanding of how the two domains interact and by offering practical guidance for implementing MLOps processes from a RE perspective, this study advances both the MLOps and RE fields. Quality feedback was collected on the artifact in the form of theoretical evaluations. However, the main shortcoming of the study is the lack of evaluation of the artifact's effectiveness under real-world conditions. Therefore, a recommendation for further research is to conduct case studies testing the artifact in real-world settings to evaluate its effectiveness and improve upon its limitations.

Keywords: Machine learning operations, Machine learning, Requirements engineering, ML, RE, MLOps, Design science research.

Acknowledgements

First and foremost, we would like to express our deepest gratitude to our supervisor, Jennifer Horkoff, for her dedication to guiding and supporting us throughout this thesis. Whenever an impasse was reached, we knew she and her expertise could be relied upon.

Special thanks to our company supervisor, Emanuella Wallin, for helping us establish contact with several experts in the field and giving us unwavering support. We are also grateful to Polestar for providing us with the opportunity to conduct our research and for all the resources they provided.

Finally, we would like to say that this endeavor would not have been possible without all the participants in this study. Their generosity to share their experience and knowledge has been invaluable to this thesis and enabled it to succeed. Without all of you, it would not have been possible, thank you for your time and contribution.

Milos Bastajic, Jonatan Boman Karinen, Gothenburg, June 2023

Contents

List of Figures **xi**

List of Tables **xiii**

1 Introduction **1**
1.1 Problem statement . . . 2
1.2 Research questions . . . 2
1.3 The purpose of the study . . . 3

2 Related works **5**
2.1 Machine learning . . . 5
2.2 Development operations (DevOps) . . . 5
2.3 Machine learning operations (MLOps) . . . 6
2.4 Requirements engineering . . . 7
2.5 RE for ML . . . 7

3 Methodology **9**
3.1 Design Science Research . . . 9
3.2 Collaborating company context . . . 11
3.3 Data collection . . . 11
3.3.1 Literature analysis . . . 13
3.3.2 Interviews . . . 13
3.3.3 Workshop . . . 14
3.4 Data Analysis . . . 14

4 Artifact **17**
4.1 The MLOps Requirements Form . . . 17
4.1.1 How to use it in practice . . . 20

5 Results **21**
5.1 Cycle I Findings . . . 21
5.1.1 Problem investigation . . . 21
5.1.1.1 Challenges found in the literature . . . 21
5.1.1.2 Best practices found in the literature . . . 23
5.1.2 Solution candidates . . . 24
5.1.3 Evaluation . . . 25
5.2 Cycle II findings . . . 27

5.2.1 Problem investigation . . . 27
5.2.1.1 Project Scoping . . . 28
5.2.1.2 Current Status of MLOps . . . 31
5.2.1.3 Data . . . 32
5.2.1.4 Infrastructure . . . 35
5.2.1.5 System Monitoring . . . 38
5.2.1.6 Results From the Pipeline-focused Interview . . . 39
5.2.2 Solution candidates . . . 39
5.2.3 Evaluation . . . 40
5.3 Cycle III Findings . . . 44
5.3.1 Problem investigation . . . 44
5.3.1.1 Confirmed themes . . . 44
5.3.1.2 Developing a Model . . . 45
5.3.1.3 Requirement Management . . . 46
5.3.2 Solution candidates . . . 46
5.3.3 Evaluation . . . 46

6 Discussion **49**
6.1 Threats to validity . . . 53
6.2 Future work . . . 54

7 Conclusion **55**

Bibliography **57**

A Appendix 1 **I**

B Appendix 2 **VII**

C Appendix 3 **XIII**

D Appendix 4 **XV**

List of Figures

3.1 An overview of the methodology utilized throughout the three DSR iterations in this study. Moreover, the figure presents how the data and knowledge from one DSR stage were fed into the next stage. Lastly, the columns show how each event relates to the DSR stages and the RQs. 10
3.2 Overview of the iterative workflow employed in design science research. Image adapted from Hevner [22]. 10

4.1 Overview of the stages of MLOps and its iterative nature. 17
4.2 Part one of the final artifact, includes requirement questions regarding the scoping stage of an ML system. 18
4.3 Part two of the final artifact, includes requirement questions regarding the data stage of an ML system. 18
4.4 Part three of the final artifact, includes requirement questions regarding the modeling stage of an ML system. 19
4.5 Part four of the final artifact, includes requirement questions regarding the deployment stage of an ML system. 19

5.1 Fishbone diagram presenting the identified themes during the thematic analysis of the first semi-structured interviews. The themes relate to the best practices and challenges found in an MLOps process. 28

A.1 First version of the artifact developed based on the literature review findings from cycle one's problem investigation. II
A.2 Front page of the second version artifact created. Supposed to serve as an introduction to the MLOps Requirements Form (the artifact). . III
A.3 Part one of artifact version two, see Figure A.4 for the second part. This is the form without the front page, the front page can be found in Appendix A. This artifact is the result of cycle two's problem investigation and the evaluation from cycle one. IV
A.4 Part two of artifact version two, see Figure A.3 for the first part. This is the form without the front page, the front page can be found in Appendix A. This artifact is the result of cycle two's problem investigation and the evaluation from cycle one. V
A.5 Front page of the final artifact created. Supposed to serve as an introduction to the MLOps Requirements Form (the artifact). VI

B.1 The script used for the first set of semi-structured interviews. VIII
B.2 The script used for the extra interview held during iteration 2 which focused on the participants' team's current data pipelines and workflows. IX
B.3 Description of the imaginary case given to the participants during the workshop. IX
B.4 Requirements sheet for the imaginary case given to the participants during the workshop. This was used as a summary of the case during the workshop. X
B.5 Set of evaluating questions given to the participants after the workshop. XI

D.1 Workshop participants Requirements Questions Answers regarding the scoping stage of the imaginary case. XV
D.2 Workshop participants Requirements Questions Answers regarding the data stage of the imaginary case. XVI
D.3 Workshop participants Requirements Questions Answers regarding the modeling stage of the imaginary case. XVI
D.4 Workshop participants Requirements Questions Answers regarding the deployment stage of the imaginary case. XVII

List of Tables

3.1 Interviewee participant traceability matrix, including their reference ID (the first number is the ID and the second number indicates that the interviewee participated in several rounds of interviews), current role, and current company. Numbers replace their name and company name as an effort of anonymization. *Minimum amount of possible experience, used where it was unclear. 12

5.1 Traceability matrix presenting which Requirement Question is derived from which Best Practice or Problem identified in the surrounding literature. Additionally, the Source column documents the origin of each Best Practice or Problem. Lastly, the rows are divided into the same MLOps stages found in the artifact, featured in Chapter 4. . 24
5.2 Traceability matrix presenting the changes made to the artifact based on the first cycle evaluation and second cycle problem investigation and which interviewee contributed to a specific change. 40
5.3 Traceability matrix presenting the changes made to the artifact based on the second cycle evaluation and third cycle problem investigation and which interviewee contributed to a specific change. 46

C.1 Codebook displaying the collection of inductive and deductive codes and their description used during the thematic analysis. XIV

1

Introduction

The use of machine learning (ML) has increased significantly in recent years and organizations utilizing ML simultaneously increase earnings while decreasing spending [1]. In order to fully leverage the benefits of ML in industry, it is critical to have a well-organized and efficient approach for how to operationalizing ML models [1]. In the context of this study, ML operations (MLOps) include the end-to-end conceptualization, implementation, monitoring, deployment, and scalability of ML products [2]. Although studies show the economic benefits seen when operationalizing ML models, research indicates that the industry struggles with taking the models to production [2]–[4].

Since MLOps is a relatively new topic, studies have been conducted to address the ambiguity of the term [2], [3], and to provide practitioners with suggestions of tools and architectures for implementing an MLOps process [2], [4], [5]. However, there exists a gap in research regarding MLOps grounded in requirements engineering (RE). To fully leverage the benefits of ML in the industry, this study presupposes that it would be beneficial for practitioners to take into consideration the RE aspects of MLOps.

To fill this gap in research, this thesis undertakes a Design Science Research (DSR) approach to investigate a set of research questions concerning RE for MLOps, described in section 1.2. Additionally, an artifact is developed in parallel to help organizations curate requirements for their MLOps. The artifact includes specific Requirement Questions, designated roles to ask these questions, and common responses to these Requirement Questions, further elaborated in Chapter 4.

In traditional software development, RE is considered to be a crucial part that ensures software meets the needs of its users and stakeholders [6], and involves identifying, documenting, and managing system or product requirements [7]. This study assumes that the introduction of RE to MLOps will give practitioners similar advantages to those achieved through the use of RE in software development.

This study was conducted in collaboration with the charging and energy division of the automotive company Polestar, which is well-positioned to provide insight into the challenges and opportunities related to MLOps and RE in the automotive

industry. Polestar is a leading player in the electric vehicle industry, known for its inventiveness and forward-thinking approach to utilizing ML in its vehicles and improving customer safety and customer experiences.

The remaining thesis is structured as follows: First, the problem statement, research questions, and purpose of this thesis are presented concisely. Second, the related work and background information are presented. Third, the methodologies utilized in this study are outlined: DSR, literature review, interviews, workshops, and data analysis methods. Fourth, the results consist of the developed artifact and the findings regarding the research questions. Lastly, this thesis is concluded with a discussion of the results, future research suggestions, and some conclusions are drawn.

1.1 Problem statement

The implementation of ML in the automotive industry can bring about numerous benefits [8]–[10]. However, multiple organizations may encounter difficulties when utilizing ML in production [11]. These challenges include: Meeting the requirements for efficient and business value-gaining models, monitoring model performance and accuracy, and the need for data validation and preprocessing [12]. To overcome these obstacles and ensure the successful implementation of ML, a structured and efficient approach, such as MLOps, is required to manage the development, deployment, and maintenance of ML models.

Based on this thesis' literature analysis, to the best of our knowledge, there has been no research published yet on how to do requirement-grounded MLOps. Research has been conducted on RE for ML [13], [14], however, since requirements for individual ML models differ from the requirements for ML pipelines and the overarching MLOps process, there is a need for more research in this field. As an example, the differences in requirements for ML models and MLOps processes could be: For ML models, it is necessary to address what data will be used for training the model. For MLOps processes, instead, one might be more interested in defining how often it is necessary to upload the training data to the ML or data pipeline.

1.2 Research questions

The study aims to answer three research questions (RQs), each with its own individual primary area of focus, the problem statement, the solution to the problem, and the evaluation of the solution:

RQ1: What are current challenges in designing an MLOps process and how do they relate to requirements knowledge?

This first research question aims to investigate the current challenges associated with designing an MLOps process and how they relate to requirements knowledge.

RQ2: Which potential solution exists to mitigate the challenges of developing an MLOps process grounded in requirements engineering?

This second research question builds upon the first research question by identifying what potential solution may exist to mitigate the challenges associated with developing an MLOps process that is grounded in RE.

RQ3: How well does the potential solution mitigate the requirements-related problems with developing an MLOps process?

This third research question evaluates the effectiveness of the potential solution identified in the previous research question in mitigating the requirements-related problems associated with developing an MLOps process.

1.3 The purpose of the study

The purpose of this study is to enhance the current understanding of the intersection between MLOps and RE, with the ultimate goal of creating an artifact to guide MLOps implementation from an RE perspective. To achieve this objective, the study will investigate the current practices of implementing MLOps processes. Additionally, the study will investigate the current best practices and challenges associated with developing an MLOps process and the relationship of these challenges to requirements knowledge, the potential solutions that exist to address these challenges, and the effectiveness of these solutions in mitigating the identified challenges. Through this artifact, the aim is to develop a strategy for incorporating requirements engineering into MLOps processes, thus achieving a more systematic and reliable approach to managing ML models in production.

2

Related works

This chapter will cover thoroughly selected information that is relevant to this thesis by addressing the following topics: ML, DevOps, MLOps, RE, and RE for MLOps. The collected information consists primarily of academic literature. However, since MLOps is a relatively new topic area there is a lack of peer-reviewed literature. Therefore, some information was sourced from specialty courses from reputable experts, not yet peer-reviewed papers, and leading industry blogs. As a result of these sources' origins, extra precaution was taken when selecting them and they were used only when no peer-reviewed literature existed that could substitute them. Lastly, where these sources are utilized in this thesis, the nature of the source is disclosed clearly.

2.1 Machine learning

Multiple aspects of ML and its relationship to the automotive industry have been studied, as demonstrated by several examples in the literature. For instance, Fernández-López et al. [8] discuss the use of ML for manufacturing optimizations in the automotive industry, while Theissler et al. [9] examine the potential for AI-powered vehicle maintenance predictions to drive cost savings. Additionally, advanced driver-assistance systems (ADAS) that utilize complex ML models are now commonly implemented as a safety feature in newer vehicles [10]. As the use of ML in the automotive industry and other industries continues to grow and become more complex, the need for effective ML operations also increases.

2.2 Development operations (DevOps)

In their paper, Subramanya et al. [15] discuss how the emergence of the DevOps approach was motivated by the need to improve the efficiency of cross-functional teams in releasing software. Traditional approaches that involved little collaboration between software developers and operations teams often resulted in difficulties in achieving smooth rollouts. The authors state that DevOps aims to address these challenges by implementing processes that enable faster, more reliable, and repeatable software builds. Additionally, automated testing and releases of software builds are further solutions for these challenges.

As the adoption of DevOps has proven successful in various organizations [16], there has been a growing interest in the practice of MLOps, which aims to bring data scientists and operations teams together in order to achieve similar benefits [17]. Furthermore, the authors acknowledge that MLOps still is in its infancy and therefore conduct a systematic literature review and a grey literature review. As a result, the authors manage to derive and validate a framework that identifies the activities involved in the adoption of MLOps, together with a maturity model for mapping the stages companies pass during the implementation of MLOps processes.

2.3 Machine learning operations (MLOps)

The goal of both MLOps and DevOps is to improve software development and deployment processes. While overlap exists between these practices, they differ in which domain they focus on. MLOps focuses on the management of an ML model's life-cycle and can be seen as an extension of DevOps which instead focuses on regular software [18]. MLOps involves integrating an ML model into a broader software development and deployment pipeline and consists of model development, training, deployment, and monitoring.

MLOps has become a trending topic in organizations looking to implement the practice, as well as in the research community. In a paper by Tamburri [3], the researcher discusses the challenges and limitations in the field of AI software operations, specifically in regard to MLOps. Tamburri identifies that while MLOps involves the orchestration of various software components to support the end-to-end life-cycle of ML models, the complexity of these operations can make them unsustainable. Tamburri argues that both research and practice in the field have focused on producing tools and components to support the definition and operation of AI software, but have not adequately addressed the sustainability of these operations. To address this issue, Tamburri proposes a conceptualization of sustainable MLOps and presents a research roadmap for its pursuit.

In an MLOps specialization course given by Andrew Ng and DeepLearningAI [19] the complete end-to-end MLOps life-cycle is covered and discussed. The course brings forth multiple challenges found within each of the MLOps stages: Scoping, Data, Modeling, and Deployment. These challenges are relevant for this thesis in such a sense that they are stage-specific and often paired together with best practices in the industry, making it possible to extract challenges, solutions, and relations to requirement knowledge. In another grey literature published as a blog post by Microsoft [20] they discuss a specific case where an MLOps process was implemented. In addition to discussing the selected tools and technologies for their MLOps process, the post discloses some of the challenges encountered when setting up the MLOps process as well as requirements for the MLOps infrastructure. These challenges and requirements were found relevant for this thesis as they informed the development of the artifact.

2.4 Requirements engineering

The authors, Nuseibeh and Easterbrook, discuss the process of RE in conventional software development [7], which involves identifying the stakeholders and their needs and documenting them in a way that can be analyzed, communicated, and implemented. The definition of RE provided in the article emphasizes the importance of "real-world goals" that motivate the development of a software system, the need for "precise specifications" that provide a basis for analysis and validation, and the fact that specifications may evolve over time and across different software systems. RE is an essential part of the engineering process as it anchors development activities to a specific problem and allows for the analysis of the appropriateness and cost-effectiveness of the solution. Additionally, the authors point out that RE plays a role in the management of change in software development, as requirements may change during development and evolve after a system has been in operation. However, most RE efforts occur early in the project, as it is more expensive to fix requirements errors later in the project life-cycle. The relationship between RE and MLOps is that while RE focuses on identifying and documenting stakeholders' needs, MLOps facilitate the efficient and reliable development and delivery of ML systems. Both areas emphasize collaboration and an iterative workflow to ensure that the developed systems meet the transforming requirements of their users and stakeholders.

2.5 RE for ML

The publication landscape of RE for ML-based systems was examined in a systematic mapping study by Villamizar et al. [14]. The authors found that the challenges of RE are exacerbated by the unique characteristics of ML. These challenges include a lack of validated RE techniques specifically designed for ML systems, an incomplete understanding of non-functional requirements (NFRs) for ML, and difficulties in managing customer expectations. The authors believe that to address these challenges and ensure the quality of ML-based systems, it is necessary to identify best practices and approaches for RE in the context of ML-based systems. This will allow for the development of appropriate and cost-effective solutions that can effectively manage change and anchor development activities to specific problems.

The field of MLOps is relatively new and there exists no set definition. In the not-yet peer-reviewed paper by Kreuzberger et al. [2], they explore the concept of MLOps and discuss how it can address the challenges of automating and operationalizing ML products. Their aim is to provide researchers and practitioners with a set of designated technologies to assist in these challenges. The paper employs a mixed-method research approach, including a literature review, a tool review, and expert interviews, to provide an overview of the necessary principles, components, roles, architecture, and workflows associated with MLOps. Additionally, the paper provides a definition of MLOps and highlights open challenges in the field. Kreuzberger et al. found 9 principles which they define as "a guide to how things should be realized in MLOps and is closely related to the term 'best practices' from the professional sec-

tor", 9 technical components needed to fulfill the 9 principles, 7 roles to implement the technical components, and 5 steps combining the previous findings to implement an end-to-end MLOps process. Furthermore, they present open challenges related to the organizational mindset of model-driven ML instead of a product-oriented approach, MLOps processes designed to match fluctuating demand, and the operational need for automation in order to handle a constant stream of data and version control of data, model, and code. This thesis builds upon this paper by utilizing the roles presented in the result to specify who to ask a specific requirements question to in our artifact. Lastly, the challenge of moving from an ML model-driven mindset to a more product-oriented approach is addressed by making the planning and implementation of an MLOps process more grounded in requirements.

3

Methodology

This chapter presents the methodology utilized in this thesis: The research approach, collaborating company, context, data gathering methods, and lastly, the data analysis approach. The section concerning data gathering consists of multiple sub-sections, each describing different methods used while gathering data for this thesis.

3.1 Design Science Research

This study follows a Design Science Research (DSR) approach. Design science is a research method that involves the development and evaluation of artifacts (such as models, theories, and prototypes) that have the potential to solve practical problems and advance scientific understanding [21]. It is often used to address complex real-world problems that cannot be fully understood or resolved through pure traditional research methods.

Knauss [21] states that DSR typically follows a systematic, iterative process that involves identifying a problem, developing a solution (the design artifact), and evaluating the effectiveness of that solution. Figure 3.1 presents an overview of all the methods used in each iteration, where each method will be further explained in this chapter. Furthermore, the design artifact is intended to be used in real-world scenarios to address an identified problem and therefore it is usually developed with a specific audience or user group in mind.

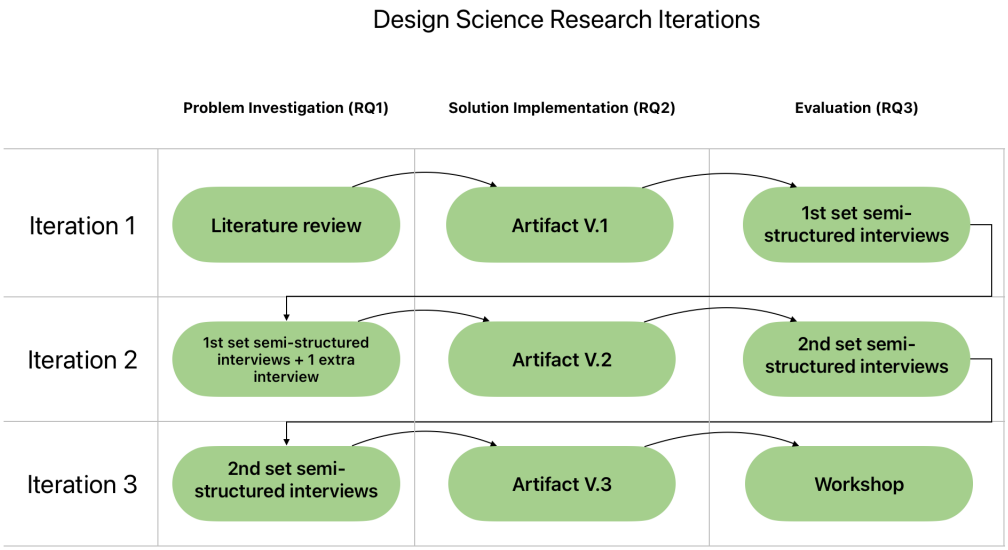


Figure 3.1: An overview of the methodology utilized throughout the three DSR iterations in this study. Moreover, the figure presents how the data and knowledge from one DSR stage were fed into the next stage. Lastly, the columns show how each event relates to the DSR stages and the RQs.

Hevner [22] argues that DSR is constructed by three parts: The Relevance cycle, the Design Cycle, and the Rigor Cycle (see Figure 3.2). The Relevance Cycle was performed in the beginning by selecting a relevant collaborating company (see section 3.2) and later iterated over in order to find additional relevant people involved. These people were then interviewees in the expert interviews held to gather further insights from the industry.

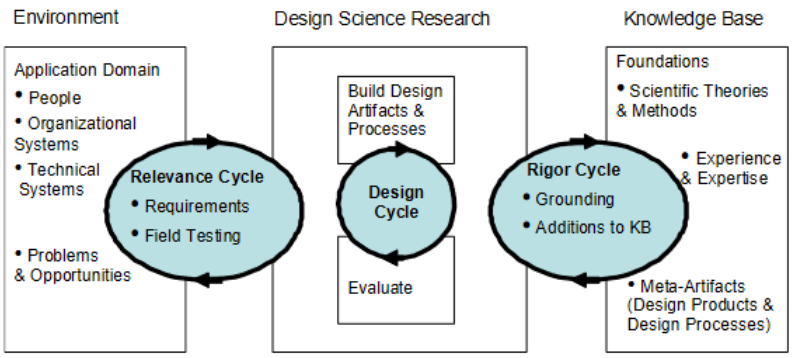


Figure 3.2: Overview of the iterative workflow employed in design science research. Image adapted from Hevner [22].

The Rigor Cycle was also mostly iterated over at the beginning of this thesis as it

was the source of knowledge for the initial artifact creation, see Figure A.1 in Appendix A. In the end we contributed to this knowledge base by organizing common challenges and best practices for MLOPs in one place. Furthermore, an MLOPs requirements form was developed which helps industry practitioners mitigate these challanges and implement best practices. Lastly, the final part of Hevner's DSR method is the Design Cycle which was iterated three times. In this thesis, the artifact was built or improved and later evaluated in each iteration in accordance with the Design Cycle.

The research questions described in section 1.2 were addressed and refined through the three iterations. Each research question had its own primary focus: The problem investigation, the solution to the problem, and the evaluation of the solution, respectively. During each iteration, all research questions were worked on, but with a primary focus on one specific research question: The first iteration primarily focused on the problem statement, while the second and third iterations mainly focused on the solution to the problem and the evaluation of the solution. This allowed for the initial creation of the artifact and its refinement in subsequent iterations, resulting in an artifact built up and refined based on data collected during each iteration.

3.2 Collaborating company context

Polestar is a global company based in Sweden that is recognized for its sustainability and innovation in the automotive industry. The company is a leading player in the electric vehicle market and is known for manufacturing premium electric vehicles. Polestar operates on a global scale, with a focus on development and production.

Polestar's charging and energy division was in the process of implementing an MLOps architecture to manage the end-to-end ML life cycle. Thus, this division was interested in exploring the potential benefits of applying requirements engineering (RE) to the design of its MLOps architecture, to ensure that the system meets their needs and aligns with industry regulations and standards.

3.3 Data collection

Multiple methods for data collection were utilized in this study: several semi-structured interviews, a literature analysis, and a workshop. These methods were selected due to their ability to provide a comprehensive and diverse range of information, facilitating a thorough understanding of the subject.

The data gathering was conducted with experts and practitioners in the field of MLOps, mostly within Polestar. However, since there was a general lack of people with experience in MLOps, additional interviews were held with experts outside of Polestar. Furthermore, it was decided against using random sampling for interviewees due to the small number of experts and instead relied on selective and snowball sampling. As a result, the sampling was carried out by first requesting possible indi-

Table 3.1: Interviewee participant traceability matrix, including their reference ID (the first number is the ID and the second number indicates that the interviewee participated in several rounds of interviews), current role, and current company. Numbers replace their name and company name as an effort of anonymization. *Minimum amount of possible experience, used where it was unclear.

Interviewee ID	**Current Role**	**Company**	**Experience (Years)**
1st set semi-structured interviews			
ID1.1	Data Analyst	Company 1	8
ID2	Software Engineer	Company 1	3*
ID3	Software Engineer	Company 1	2
ID4	Software Engineer	Company 1	1
ID5	Software Developer	Company 1	2*
ID6.1	Product Owner	Company 1	15
ID7.1	Sr Manager, ML Engineering and Research	Company 2	20*
ID8	Software Engineering Manager	Company 1	8
2st set semi-structured interviews			
ID1.2	Data Analyst	Company 1	8
ID6.2	Product Owner	Company 1	15
ID7.2	Sr. Manager, ML Engineering and Research	Company 2	20*
ID9	Sr. Data scientist	Company 3	15
ID10	ML Researcher	Company 4	5
ID11	Software Engineer	Company 1	0.5

viduals with relevant knowledge and expertise from the company supervisor. Once the initial few interviews were conducted, it was possible to ask the interviewees for recommendations on additional relevant interview prospects. For a complete interviewee overview, see Table 3.1.

In order to evaluate the first version of the artifact developed from the results of the literature analysis and simultaneously continue the problem investigation, the semi-structured interviews were divided into two parts: First, the interviewees were asked questions related to the problem investigation for the next cycle. Second, at the end of the interviews, they were asked questions targeting the evaluation of the artifact created from the current cycle's problem investigation. The interview structure of combining problem investigation together with evaluation of the previous artifact was used for all general interviews. For a complete list of all interview questions, refer to Appendix B.

3.3.1 Literature analysis

A literature analysis was conducted at the start of the project to acquire a comprehensive understanding of existing research and knowledge related to RQ1. The purpose was to identify gaps in knowledge, build on existing research, situate the work within the broader context of the field, and ensure that the initial proposed solution was based on previous findings and theories, following the workflow visualized in Figure 3.2. The results of the literature analysis guided the development of the first artifact, see Figure A.1 in Appendix A. Identifying common challenges and best practices found in the surrounding literature made it possible to create a comprehensive set of Requirement Questions that served as a foundation for future Design Cycles. This approach was considered necessary due to the low number of available people in the industry with adequate knowledge of MLOps and its related challenges.

The literature analysis was performed using literature database search engines Google Scholar and Web of Science, which searched all established and relevant journals for this study, such as IEEE and ScienceDirect. To begin the search, terms such as: "RE for MLOps", "RE for ML", "MLOps challanges", "DevOps for ML", and "MLOps best practices" where searched. After a few relevant papers had been found a snowballing approach was utilized to find further relevant studies. As previously mentioned, grey literature was selected with great precaution and was only used when peer-reviewed resources were lacking, making up a minority of cited sources.

3.3.2 Interviews

Interviews were held during all cycles of the study and followed a semi-structured format, see Table 3.1 for a full list of interviewees. Doody and Noonan [23] define semi-structured interviews as a common method used in qualitative research. They involve using predetermined questions, but the researcher is also free to seek clarification and explore issues that arise spontaneously. Thus, interview protocols were developed to collect similar types of data from participants and create a sense of order. The questions were designed to be flexible and open-ended, in order to be able to vary the sequencing and wording of the questions and to ask additional questions. This allowed for the exploration of new paths that emerged during the interviews that may not have been considered initially. Additionally, a pilot interview was held in each iteration to test and improve the interview protocol before conducting the remaining interviews. These interviews were conducted via video conference and they were recorded and automatically transcribed using Microsoft's Teams application for later analysis. As a precautionary measure, both researchers analyzed the recordings afterward and corrected any possible mistakes in the automatically generated transcript.

As seen in Figure 3.1, the first set of semi-structured interviews was supported by an additional targeted interview with one of the participants that were in the first set of interviews. This additional interview focused on the interviewee's current data pipeline and ML workflow, see Figure B.2 in Appendix B for the interview script.

The conducted interview provided valuable insights into the current processes and challenges faced by the team, which contributed to the development of the artifact.

3.3.3 Workshop

Ørngreen and Levinsen [24] categorize workshops into three categories: workshops as a means, workshops as practice, and workshops as a research methodology. The latter was utilized in our study and the workshop was according to the authors' description "... specifically designed to fulfil a research purpose: to produce reliable and valid data about the domain in question [24]". The workshop was conducted in the final stage of this study, its goal being to evaluate the final artifact developed through the DSR process. Workshops were not utilized in the earlier stages as it was believed that they may be more time-consuming and less beneficial without a more finalized artifact.

The conducted workshop included three participants who had previously partaken in the interviews: ID1, ID6, and ID7. These individuals were purposely selected as they could cover a few of the roles in the artifact, which was necessary due to there only being three participants (see roles in Chapter 4). The workshop aimed to evaluate the final artifact developed in this thesis and was arranged in three steps: Firstly, an imaginary business case involving MLOps was presented to the participants, who role-played as the team who would implement the solution to the presented business case. Secondly, the group got to apply and use the artifact as a tool to come up with a project plan, which consisted of a multitude of MLOps requirements. While the group worked with the case and the artifact, the hosts noted any apparent struggles and interesting discussions with time stamps, which were later analyzed using the recordings from the workshop. Lastly, the workshop participants answered a set of evaluating questions individually. The aim of these questions was to capture the participant's thoughts on working with the artifact on a staged real-life business case. Refer to Appendix B to see the evaluating questions and case description used in the workshop.

3.4 Data Analysis

The data collected in this study consisted solely of qualitative data, which was analyzed using coding and theme identification methods presented by Saldaña [25]. Coding is a method of analyzing qualitative data by identifying patterns and themes within the data and assigning codes to these patterns. All final codes can be seen in the codebook in Appendix C. While reading through the transcriptions of the interviews open coding, also referred to as initial coding, was used to assign codes to each piece of data that related to the research questions. This was followed by axial coding, where the relationships between the codes were analyzed to identify higher-level themes. This made it possible to identify trends and patterns in the data and gain a deeper understanding of the topics.

In addition to the initial coding and axial coding, what Saldaña calls "Themeing

the Data" was used to find overarching themes from the collected data corpus. This involved identifying common themes that emerge across multiple sources of data and grouping the data according to these themes. This made it possible to see how the different data sources relate to one another and identify any discrepancies or inconsistencies, as well as enabled the researchers to find a coherent narrative.

To ensure the reliability and validity of the data analysis, several strategies were employed. Firstly, both authors of this thesis independently analyzed the data. Secondly, the results of the analysis were compared to ensure a consistent interpretation of the data. Moreover, where possible, data was triangulated by collecting it from multiple sources, such as interviews, literature, and workshops, to ensure results were not biased by a single data source.

4

Artifact

The following chapter focuses on explaining the artifact that was designed through three Design Cycles to address the challenges identified as results of the RQ1: *What are current challenges in designing an MLOps process and how do they relate to requirements knowledge?*. The artifact acts as the solution and results of RQ2: *Which potential solution exists to mitigate the challenges of developing an MLOps process grounded in requirements engineering?*. This chapter presents the design and the idea behind the artifact, which is an MLOps Requirements Form, and gives suggestions on ways of using it. The rationale for the artifact design is explained with the results in Chapter 5.

4.1 The MLOps Requirements Form

The MLOps Requirements Form is a tool that is developed in order to assist teams or individuals in eliciting MLOps requirements when implementing MLOps. The artifact is structured to align with the end-to-end stages of an MLOps process: Scoping, Data, Modeling, and Deployment. Figure 4.1 gives an overview of the stages and a visual representation of how information from one stage feeds into another.

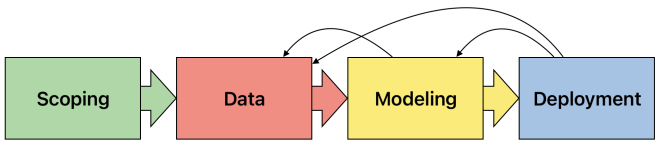


Figure 4.1: Overview of the stages of MLOps and its iterative nature.

The MLOps Requirements Form consists of two parts: The first part serves as an introduction that guides users on how to use the form effectively, see Figure A.5 in Appendix A. The second part is the form itself, which is the result of cycle three's

problem investigation and the evaluation from cycle two, see Figure 4.2, 4.3, 4.4, and 4.5.

Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Scoping:				
	Business stakeholder	What are the business problems?		Battery optimization, Fruad detection, Demand forecasting
	Data Scientist	Can the business problems be solved with ML, how?		Has it been done before, research proves it possible, still unclear
	Product owner	What are the metric for success?		ROI, customer wishes
	Product owner	What are the resources needed?		Data, time, people
	Product owner, Business stakeholder, Data scientist	What is the budget limit for the computation necessary to train the model?		If on premise: 100h allowed, 50h, Unlimited If on cloud: Budget is $1,000, $5,000, $500
	Business stakeholder	Who is the end user?		Demographical information, Internal company users, Customers
	Business stakeholder	How will the users interact with the model, what interface will they need?		App, Voice activated feature, Web page, API
	Business stakeholder, Product owner, Data scientist, Data engineer	Who is the domain expert and can we access them?		Doctors, Lawyers, Domain specific researcher

Figure 4.2: Part one of the final artifact, includes requirement questions regarding the scoping stage of an ML system.

Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Data:				
	Business stakeholder, Product owner, Data scientist, Data engineer	Where does the data come from?		Owned data, crowdsourced, purchase data, purchase labels
	Data scientist, Data engineer	What data format will be used?		Structured, unstructured
	Data scientist, Data engineer	How should the data be preprocessed?		Remove data, remove duplicates
	Data scientist, Data engineer, Domain expert	What are the data labeling guidelines?		On images: Label each scratch independently on the screen, label each animal separately in the field
	Product owner, Business stakeholder	Who will label the data?		In-house resources, Crowdsourced, Outsourced, Mixture of resources
	Data scientist, Data engineer, Product owner	What meta-data should be collected?		Time, system model, factory, device type
	Data engineer, Legal team, Business stakeholder, Product owner	Are there any privacy concerns regarding the data?		Names, Emails, Addresses, Phone number, general GDPR concerns
	Data engineer, Legal team, Business stakeholder, Product owner	Are there any necessary data ownership considerations?		Data is owned by us, it's open source, another party owns all data
	Product owner, Data scientist, Data engineer	How much data is expected to be stored?		~10TB
	Product owner, Data scientist, Data engineer, Domain expert	When does the data become irrelevant?		Never, new product version release, annually
	Data engineer, Domain expert	Are there any cyclic behaviours to the data?		Seasonal sales cycle, full day cycle
	Data scientist	What is the minimum amount of data that is necessary to train the model?		10k images, 100 gb worth of 1080p mp3 video recordings
	Data scientist	For streaming data, what is the minimum frequency of data points necessary to meet the business goals?		Every 5ms, Every 1s, Every data point
	Product owner, Data scientist, Data Engineer	How will the data be acquired?		Automated tool, manually collected, purchased

Figure 4.3: Part two of the final artifact, includes requirement questions regarding the data stage of an ML system.

Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Modeling:				
	Product owner, Data scientist	What is the model baseline?		Human-level performance, A previous system's performance, Dummy model
	Product owner, Legal	Is it necessary to audit the model? Who should audit the model? What is the audit focus?		Yes/No. Business stakeholder, Third party, Data scientists. Transparency, Equality, Fairness, and Accountability...
	Data scientist, Data engineer	Which potential risks for bias exists?		Gender bias, Brand bias, Ethnicity bias
	Product owner, Data scientist	How is the input data served to the model?		Batch data, Real time data
	Data scientist, IT Architect	Where should the experimental data result be stored?		Database, Excel document, JSON-file
	Product owner	What are important business goal metrics the ML model should consider?		Business required classifications performance, different from general ML model performance
	Data scientist	What experimental data should be tracked?		Dataset used, Hyperparameters, Results, Results with metric summary/analysis, Training resources, Training time),
	Data scientist, Software engineer, DevOps engineer, MLOps engineer	What deployment constraints exist?		None, Edge device's hardware capabilities

Figure 4.4: Part three of the final artifact, includes requirement questions regarding the modeling stage of an ML system.

Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Deployment:				
	Product owner, MLOps engineer, DevOps engineer	How should the deployment process be handled?		Canary releases, A/B releases, Shadow releases
	Product owner, MLOps engineer, DevOps engineer	Where should the prediction device be located?		Cloud or edge device
	DevOps engineer, MLOps engineer, Software engineer	Which software metrics are important to monitor?		Memory, computing power, latency, throughput, server load
	Data scientist, Data engineer, MLOps engineer	Which input metrics are important to monitor?		feature types (INT or String), feature range, Data schema validation
	Data scientist, Software engineer, MLOps engineer	Which output metrics are important to monitor?		# times users redo search, avg. prediction accuracy
	Product owner, MLOps engineer, Data scientist	How often should the model be retrained on the data gathered from deployment?		Every Monday, once a month, based on deployed input/output metric triggers
	Product owner, DevOps engineer, Data scientist	Are there any specific performance requirements?		Latency requirements, Query per seconds requirements

Figure 4.5: Part four of the final artifact, includes requirement questions regarding the deployment stage of an ML system.

The Requirement Questions in the MLOps Requirements Form are sorted mainly by which stage they are most related to. However, the order of the questions in the different stages was also considered in order for them to mirror how these questions might come up naturally in an ML project which needs to be operationalized.

Each Requirement Question in the form is intended to be answered by specified roles within a team and sometimes roles outside of the implementation team. Additionally, each question includes a field for documenting the answer, paired together with some example answers which can be used for reference or clarification of said Requirement Question. The output of the form are the documented Requirement

Question Answers, which can be interpreted as informal requirements for MLOps requirements. These requirements can then be used as they are, or as a foundation for creating more formal requirements.

The MLOps Requirements Form was intentionally designed to be general, making it applicable to a wide range of industries and projects. Based on an extensive literature analysis, these questions have been informed by best practices and common challenges that arise when implementing MLOps processes. Through interviews with practitioners who had practical experience with MLOps, or relevant areas, the questions in the form were refined and improved iteratively. Each question in the form has been designed to serve as an adaption or mitigation to one or more of the best practices and challenges identified during the literature analysis and interviews.

The MLOps Requirements Form draws on the collective wisdom of the surrounding literature and MLOps experts interviewed, providing practitioners with a comprehensive and flexible approach to elicit and document requirements for an MLOps process. This approach ensures that those who use the artifact have considered many essential factors and requirements, regardless of their specific project or industry, thereby reducing the risk of common challenges derailing their MLOps implementation efforts.

4.1.1 How to use it in practice

How one chooses to use the artifact can depend on their background, the project size, and company structure. As an example, in the scoping phase of a project, a meeting could be arranged with all of the relevant roles where all questions could be discussed and answered. Alternatively, a designated person could ask each Requirement Question in a one-on-one setting.

The artifact's purpose is to guide relevant roles to discuss and answer specific MLOps Requirements Questions, regardless of how one chooses to use the artifact in practice. These answers, which are either formal or informal requirements, can then be shared with relevant parties such as the implementation team or stakeholders. As the answers can be interpreted as informal requirements, they can be used as the foundation for creating more formal requirements if that is necessary for one's specific case.

5

Results

The following chapter presents the collective results found during this thesis work. The chapter is divided into three sections, one for each cycle, also referred to as iteration. Each cycle includes a subsection for the Problem investigation (RQ1), a Solution candidate (RQ2), and an Evaluation of the Solution candidate (RQ3).

5.1 Cycle I Findings

The initial challenges, best practices, and requirements for MLOps were extracted during the literature analysis. While the initial challenges and best practices found are introduced in subsection 5.1.1, a multitude of Requirement Questions that were derived from these results are presented in subsection 5.1.2.

5.1.1 Problem investigation

This subsection presents the challenges and best practices discovered through the literature analysis, chosen because they were general and applicable to a broad range of MLOps cases. Based on these discoveries, a series of questions (see subsection 5.1.2) were formulated to elicit information for overcoming challenges and adopting best practices. The responses to these questions may serve as an initial set of requirements for an MLOps process, which can then be improved upon over time.

5.1.1.1 Challenges found in the literature

P1 - Data Drift: Data drift refers to the challenge where the distribution of input data changes from the data used to train the model, but the desired prediction output remains the same. This shift in data can cause the model's prediction accuracy to decrease, as it was trained on a different distribution of data. The rate of drift can be either slow or fast. To combat this problem, monitoring can be implemented to track changes in the input and output distribution. When drift is detected, the model can be retrained, with appropriate updates made to ensure its accuracy and reliability [4], [5].

P2 - Concept Drift: Concept drift occurs when the functional relationship be-

tween a model's inputs and outputs changes, resulting in a modification of the output definition as the input changes. Consequently, the model's previously learned patterns become obsolete, resulting in a decline in its prediction accuracy. To overcome this issue, similar to **P1**, monitoring the input and output distributions and constant retraining are effective solutions [26], [4], [5].

P3 - Inter-team Communication: Communication challenges in MLOps can arise due to the different roles and knowledge levels of professionals involved, as highlighted by Kreuzberger et al. [2]. Ng [19] presents an example of conflict when an ML model performs well on test sets but fails to meet business goals, causing disagreement between ML and business teams. Evaluating models based only on average error rates can neglect critical examples and lead to unsuccessful deployment. Moreover, Kreuzberger et al. [2] emphasize that in order to succeed with MLOps projects, it is necessary to not rely on a single role, but rather a whole group of professionals and strive for effective communication within the teams.

P4 - Performance During Serving: Performance-related challenges after deployment are commonly related to two different categories: First, traffic management concerns, such as network latency, ML system throughput, and access points. Second, ML model performance, this category of challenges relates to issues such as true label availability of the data exposed to models used for prediction. True labels are sometimes only irregularly available which makes it challenging to monitor a deployed model's current performance [26]. This stage of the MLOps process is often referred to as serving and the presented challenges were found or discussed in 13 sources in a literature review on MLOps [4] and Ng's expert course [19].

P5 - Disorganized Data: The data gathered for a model might originate from different sources and is usually disorganized. Therefore, it usually is difficult to use this raw data as input for the model, since it can not process data of this form [20], [19], [2].

P6 - Sustainable MLOps: In their paper, Tamburri et al. [3] delve into the development and implementation of sustainable MLOps and highlight three critical components: Explainability, fairness, and accountability. They first emphasize the importance of explainability, which involves providing comprehensible explanations for the logic behind automated decision-making. They then discuss fairness and the need for ML systems to allocate decision-making power fairly to all stakeholders while mitigating any biases or discrimination. Finally, they touch on accountability and the responsibility of correcting any misaligned features, and who should be held accountable for them. Tamburri et al. conclude by stating that sustainability itself is dependent on the implementation of these concepts, as they are interconnected and essential for the ethical deployment of ML. Essentially, explainability leads to observable and self-improving MLOps, fairness is necessary for sustaining social contracts, and accountability aligns with legal establishments. However, Villamizar et al. [14] point out that the primary challenge in ML is the lack of knowledge regarding specific non-functional requirements (NFRs), such as explainability, fairness,

and accountability. They explain that the understanding of NFRs is incomplete, and practitioners struggle to define and refine NFRs in an ML context.

5.1.1.2 Best practices found in the literature

BP1 - Versioning: Implementing versioning for data, models, experimentation logs, and code increases a system's reproducibility and traceability, as stated in [2]. Each component of the system requires a different storage component, such as a model registry for models, a feature store for features, a pipeline store for data and ML pipelines, a regular source code repository for ML and IaC scripts, and storage for metadata, including hyper-parameters and model metrics [5]. This enables organizations to maintain a comprehensive record of any changes made to their systems, making it easier to identify errors or bugs introduced during development. Additionally, versioning empowers data scientists to confidently reproduce results, validate findings, and build on past work more efficiently.

BP2 - Model Deployment and Serving: Kumara et al. [5] present "Model Deployment and Serving" as one out of eight categories of requirements for MLOps environments. In this category, the authors discuss the importance of defining what model prediction serving pattern to use: Model-as-Service where your model is exposed as an endpoint on the web, Precompute where your model expects batches of input data to do predictions that are saved for later, or Model-as-Dependency where your model is loaded in real-time. Moreover, the authors explain that each of these methods may require specific architectural designs in order to work e.g. depending on online or offline serving or if input data consists of real-time or batch data. These are only some examples that make this an important best practice to consider when setting up an MLOps architecture.

BP3 - Data Quality and Labeling: The importance of proper data quality is another factor to consider when building ML systems, states Vogelsand and Borg [13]. The authors communicate that it is common for practitioners to use publicly available datasets for training. However, what these datasets give in terms of availability, they lack in quality and unbiasedness. The authors state that these publicly available datasets possess these downsides due to how they were labeled. This stresses the importance of labeling processes that are coordinated and transparent in situations where data quality is of importance, which often is the case for any ML system.

BP4 - Feasibility: Vogelsang and Borg [13] points out that data scientists often make development decisions in machine learning systems without considering the business domain and stakeholder needs. Ng [19] emphasizes this in his expert course, noting that identifying the right projects for ML is a rare and valuable skill in the field. Ng suggests that exploring business problems that could be solved with ML, rather than looking for ML problems, should be the first step in the development process. By gaining a clearer understanding of the problem, it becomes easier to find solutions. Once potential solutions are proposed, their feasibility and value should be assessed before considering metrics for success and resource budgeting.

5.1.2 Solution candidates

Based on the challenges and best practices found during the initial problem investigation presented in subsection 5.1.1, a first version of the artifact was created.

The artifact which can be found in Appendix A in Figure A.1 served as the foundation for future artifacts. Table 5.1 is a traceability matrix that illustrates the connection between the Requirement Questions in the artifact to the best practices and challenges the questions were derived from.

Table 5.1: Traceability matrix presenting which Requirement Question is derived from which Best Practice or Problem identified in the surrounding literature. Additionally, the Source column documents the origin of each Best Practice or Problem. Lastly, the rows are divided into the same MLOps stages found in the artifact, featured in Chapter 4.

Problem (P):	Best Practice (BP):	Source:	Requirement Question:
Scoping:			
	BP4: Feasibility	[13], [19]	What are the business problems and can they be solved with AI?
			What are the metric for success?
			What are the resources needed?
Data:			
P5: Disorganized data		[2], [19], [20]	How should the data be preprocessed?
			Where does our data come from?
			What data format will be used?
			What meta-data should be collected?
	BP3: Data quality and labeling	[13]	What is the data labling standard?
Modeling:			
P3: Poor communication between ML and business teams		[2], [19]	What are the model baselines?
			What are important business goal metrics the ML model should consider?
P6: Sustainable MLOps		[3], [14]	Is it necessary to audit the chosen model? Who should audit said model?
			Which potential risks for bias exists?
	BP2: Model Deployment and Serving	[5]	How is the input data served to the model?
			What deployment constraints exist?
	BP1: Versioning	[2], [5]	Where should the experimental data result be stored?
			What experimental data should be tracked?
Deployment:			
P1: Data drift, P2: Concept drift		[4], [5], [26]	Which software metrics are important to monitor?
			Which input metrics are important to monitor?
			Which output metrics are important to monitor?
			How often should the model be retrained on the data gathered from deployment?
P4: Performance during serving		[13], [19]	Where should the prediction device be located?
			Are there any specific performance requirements?
			How should the deployment process be handled?

The Requirement Questions were formulated to obtain relevant information from specific roles in a team, resulting in informal requirements for an MLOps project. This approach facilitates a more streamlined process for developing requirements and ensuring alignment with best practices.

5.1.3 Evaluation

This subsection presents the findings of the evaluation phase during the first set of interviews, which aimed at collecting feedback on the usability, usefulness, and content of the artifact created based on the challenges and best practices found in the surrounding literature. While this subsection reports the collected feedback, the actions taken to address the results from the evaluation are presented in subsection 5.2.2.

Several areas requiring improvement were identified and presented in this subsection. These findings served as a guide for the changes in the next version of the artifact. Table 3.1 gives an overview of the eight individuals interviewed for the evaluation stage of cycle one.

Redundancy and Clearness

During the evaluation of the artifact, we asked the interviewees about the redundancy and clearness of the artifact. Regarding redundancy, there were no complaints, however, regarding clearness, one participant directly asked how the artifact should be used. This question made both authors realize that several people had been similarly confused but had asked the question in less direct ways. This realization highlighted the need for a way to easily relay the information of how the artifact could be utilized in an industry setting.

Artifact appreciation

When asked if the artifact would be useful and why 7/8 participants said that they think the artifact would be useful while implementing an MLOps process. One reason why it would be useful, which was echoed by several of the interviewees, was that it could manage expectations and be a helpful way to communicate what is needed in order to create an MLOps process and what can be expected from it.

> *"Yes, I think so. You would have a more structured way of scoping your projects and you would have an easier time of clearing out inconsistencies when it comes to people's expectations of different things in your product."*
> - ID4

Another reason repeated by several of the interviewees was that it is a good way to get an overview of what is needed before starting the implementation of the MLOps process. A way to eliminate any creeping issues that might become costly if not spotted early on.

Suggestions

During our initial evaluation of the artifact, we received feedback from 2/8 interviewees who suggested that improvements could be made with regard to the model's exposure to end users after deployment. Specifically, they recommended adding a

Requirement Question to address the type of interface that the end user would have with the ML system. This would help answer questions about how the model would be utilized in practice. One of the interviewees also emphasized the importance of understanding the demographics of the end user, which could help determine the type of input data that the model should be trained on. For example, if a voice recognition system was specifically designed for children, data scientists would be able to create a more accurate model by filtering the training data to reflect the characteristics of the primary user group.

In 4/8 interviews, the participants suggested that the role of the business stakeholder in the artifact should be replaced with more suitable roles, particularly in the scoping section where the business stakeholder was the only role. One interviewee noted that questions regarding the appropriateness of ML solutions for business problems are better answered by a data scientist or someone knowledgeable in ML and the specific context, rather than the business stakeholder. Two other interviewees reinforced this by stating that either an ML researcher or data scientist would be better equipped to answer questions about the feasibility of ML in problem-solving. Additionally, another interviewee suggested that the scoping section should involve input from either an MLOps or DevOps engineer. The unanimous feedback highlighted the need for a thorough review and update of the roles for the next version of the artifact.

One interviewee emphasized the significance of involving subject matter experts in the project. If the current data scientist is not a domain expert, it is important to locate someone who is and establish communication. Additionally, another interviewee corroborated this claim by asserting that a domain expert is indispensable for supervised models that require labeling. They further explained that this was particularly true when developing labeling standards and guidelines, where a domain expert should collaborate with a data scientist. The domain expert provides guidelines for what should be labeled as what, while the data scientist takes care of the technical aspects of labeling. These responses indicated the necessity of addressing domain experts in the next version of the artifact.

An interviewee suggested we split up the first Requirement Question in the artifact ("What are the business problems and can they be solved with AI?") as it was considered relevant, but too broad. This was agreed with since both "What are the business problems?" and "Can the business problems be solved with ML, how?" are big enough questions to write elaborate requirements for. Furthermore, keeping "AI" was considered slightly inaccurate and would therefore be replaced by "ML" in future artifact versions. Lastly, splitting this question into two parts enabled the artifact to more accurately target the responsible roles to answer each individual question.

The Data stage, seen in Figure A.1 in Appendix A, was deemed insufficient by 3/8 interviewees. The interviewees stressed the importance of this stage and the need for a broader range of questions. Two of the interviewees suggested adding a Require-

ment Question that would address the expected size of the data. Additionally, one of the interviewees stated that explicitly writing a requirement for the approximate minimum data size necessary would be possible. Moreover, an interviewee suggested adding questions regarding the duration for which the data would need to be stored. In addition, the question of when to discard the data and when it becomes irrelevant evidently warrants attention. Lastly, one interviewee pointed out that Requirement Questions seeking answers regarding data privacy and ownership must be included.

> *"I don't think you're talking about privacy and data ownership, but maybe it should be in the scoping. Because if you work within the same company, maybe it's not much of a problem, but if it's two different entities then it becomes quite a big problem. So like even in between the same group, you still need to be careful with how you deal with the data."* - ID1.1

Furthermore, another interviewee pressed the importance of considering if there are any cyclic behaviors regarding the data that may require monitoring. Finally, a plan for detecting and addressing errors or faults in the data could be developed. These issues will be addressed in greater detail in the subsequent sections of this thesis.

Lastly, 2/8 interviewees said they would have liked to see a number of Requirement Questions that would focus on specific infrastructure requirements, e.g. specific tools, databases, and hardware for training. This was taken into consideration but was deemed unnecessary with the motivation that an IT Architect would be able to determine the required infrastructure based on the answers given to the rest of the Requirement Questions.

5.2 Cycle II findings

This section presents the results obtained during the second DSR iteration. The structure of this section is identical to the previous cycle: First, the Problem investigation is presented, then the Solution candidates, and lastly, the data from the Evaluation is introduced.

In contrast to the previous version described in subsection 5.1.1, this iteration focused exclusively on the information obtained from the semi-structured interviews, as elaborated on in section 3.3. Although the findings from cycle two are outlined in this section, the discussion on how it relates to the challenges and best practices discovered during the literature analysis can be found in Chapter 6.

5.2.1 Problem investigation

This subsection presents the results of the thematic analysis conducted on the first part of the first set of semi-structured interviews. As thoroughly described in section 3.4, the objective of the analysis was to identify common themes and patterns in the data related to RQ1. The codes and themes that emerged from the analysis, along with their definitions and examples, are presented in this subsection and Fig-

ure 5.1. These results are later used to inform the improvement of the artifact, as described in subsection 5.2.2.

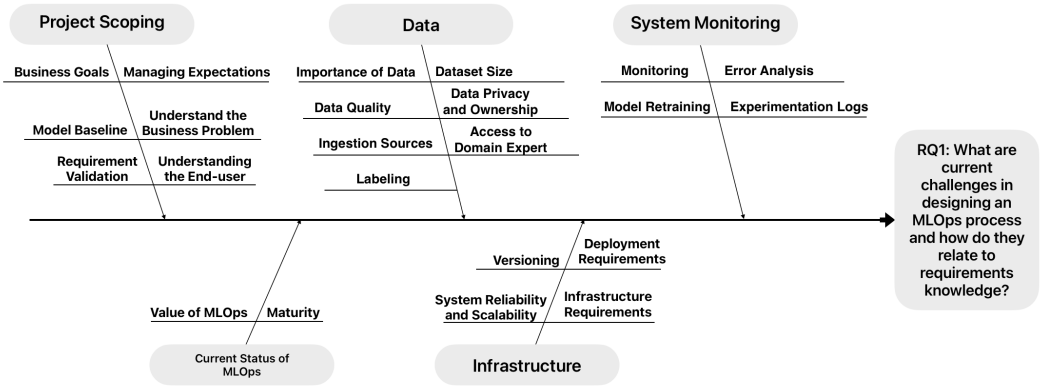


Figure 5.1: Fishbone diagram presenting the identified themes during the thematic analysis of the first semi-structured interviews. The themes relate to the best practices and challenges found in an MLOps process.

5.2.1.1 Project Scoping

The "Project Scoping" theme encompasses a set of codes related to the scoping stage of an MLOps project. This theme includes the following codes: Business Goals, Understand the Business Problem, Model Baseline, Manage Expectations, and Requirement Validation.

"Project Scoping" was discussed by 7/8 interviewees and the theme emphasizes the importance of incorporating multiple perspectives and requirements into the development process, as well as working continuously with understanding the customer and their needs in order to create successful ML products.

Business Goals: Multiple interviewees discussed the importance of incorporating various perspectives and requirements into the development process to create ML products that are deemed successful. Additionally, at least one interviewee described the challenge of capturing business requirements, incorporating them into ML research, and lastly mapping them together.

> *"That's actually a big issue with machine learning development I think (RE for ML). So some of the stuff that I personally have been focused on the last three or four years is how you capture requirements from the business side and sort of make that a part of the researcher's work. So when they do their research and they develop candidates and they do stuff, they actually can map those results right to the business requirements."*
> - ID7.1

Another interviewee brought up more general challenges with RE such as under-

standing the customer's visions and trying to deliver beyond the requirements that they can state themselves. Although this challenge is general for RE, when factoring in the complexity of ML, this challenge was found to be important to consider.

> *"... trying to understand what vision they have with the product and also ... what they think they want, but usually a product is trying to give customers also something that they don't know yet that they need. So trying to understand what their needs are and finding a way to give them that ... "* - ID6.1

Requirement Validation: To ensure effective implementation and meeting business objectives, it is crucial to thoroughly understand requirements and avoid misinterpretation or insufficient detail in their documentation.

> *"I would say the challenge is always that the intention of the requirement is understood. And not over-interpreted in the wrong way or that a requirement is misinterpreted and wrongly implemented. Or, that the requirement is written too thin so the person implementing the requirement misses a lot of things around the requirement."* - ID8

Understand the Business Problem: One challenge in ML is identifying appropriate problems for its use. Due to ML, and MLOps, being complex and resource-demanding, it is important to evaluate whether a business problem requires an ML solution or can be solved using traditional optimization. Careful assessment of the problem's nature and requirements is essential before deciding if some ML solution is optimal, followed by the decision of which ML algorithms or datasets are suitable for said problem.

> *"I see that as a part of the development, you first need to identify what type of problem you try to solve. If you should use machine learning or if you should use classical estimation. And then when you look at the data you see if you can model it etc."* - ID5

> *"One of the challenges is to understand what's an appropriate problem to be solved with machine learning. And right now there are so many areas where machine learning can be introduced that you can deliver very good results, but there are a lot of areas where you just shouldn't introduce machine learning. ... Is this, you know, a discrete function that you can code a solution or use traditional optimization? Or does this actually get better with scale as they say, and if it is fit as a machine learning candidate solution?"* - ID7.1

In conjunction with proposing ML solutions to customers, an interviewee discussed the challenges of explaining the concept behind ML and how it works for them.

> *"Customers say they want to have machine learning, they think it's in-*

> *teresting, innovative, and cool, but maybe they don't understand how it should be used and they are still stuck in a very predictive way of working. So having machine learning behind the scenes is not always that easy. How to explain (ML to the customer)?"* - ID6.1

Model Baseline: One important aspect to consider in the scoping stage of an MLOps project is what baseline the final model should have. We see that it is important to consider previous solutions and performance requirements.

> *"Usually they (business stakeholders) say give better results than the previous solution (ML solution)."* - ID1.1

However, we also see that it is sometimes important to consider other aspects such as ML versus manual labor and its significance to business value.

> *"... if you introduce machine learning in the right spot, like say you have a current problem and the only way to solve it requires a lot of manual labor and a lot of time. And maybe it's brittle and it has to be reworked every six months or something. That's the bar you have to meet with machine learning. ... if that's a 90% accurate solution during this mechanical way, discrete way, you can maybe get an 85% (with ML) ... fully automated (solution) with no manual labor and maybe even works better with scale. So actually improves with scale. And if you get to that point then you've reset the expectations on the actual quality to the way you actually produce it and the cost."* - ID7.1

Managing Expectations: Five interviewees named the importance of establishing clear expectations from the end-user of the ML system when implementing ML solutions, as there is often a knowledge gap between what ML can actually accomplish and what people assume it can do. By clarifying these expectations upfront, potential misunderstandings can be avoided, and the project can be set up for success.

> *"There's always the need to sort of get the right expectations upfront from business because you know a lot of people know what machine learning is and assume what it can do. And sometimes those assumptions aren't right so it is good To clarify that at the beginning."* - ID7.1

Understanding the End-user Interviewees emphasized the significance of identifying the end-users and their intended use of the ML solution. This can be a challenging task since it can be difficult to determine the end-user, yet it significantly impacts the nature of the final ML solution. Thus, having a comprehensive understanding of the demographics and intended use of the model by the end-users is essential for effective ML solutions.

5.2.1.2 Current Status of MLOps

The theme "Current Status of MLOps" was mentioned by 7/8 interviewees. It is a smaller theme that consists of two codes: Value of MLOps and Maturity. True to its name, the theme encapsulates the current status of MLOps, including why MLOps is necessary and why it is not used more in the industry. The findings related to this theme were deemed unrelated to RQ1 and are therefore not considered during the next Design Cycle presented in subsection 5.2.2.

Value of MLOps: Interviewees highlighted the benefits of utilizing an MLOps architecture. Specifically, one interviewee mentioned how DevOps practices are essential for developing and maintaining ML products. Since ML products involve various types of artifacts, such as models, datasets, and configurations, that need to be versioned, tested, and tracked to ensure their quality and reliability.

> *"DevOps would be really important for machine learning products because I'm seeing how we are able to organize all different software artifacts within software development. So in machine learning, I'm pretty sure there's different types of artifacts that need to be versioned, need to be tested, and need to be tracked and maintained."* - ID3

Another interviewee discussed the challenges related to bug tracking in ML projects and suggested the value a pipeline would have in minimizing the problem of bug tracking in ML systems.

> *"I would say that having this MLOps pipeline is one way to mitigate it (problems with bug tracking). But then, of course, it's very expensive. But I still think it's the right way to have it. So to enable an MLOps pipeline."* - ID6.1

Maturity: The maturity of MLOps was found to be a reoccurring topic when discussing the current state of MLOps. Three out of the eight interviewees stressed that MLOps currently is very immature, with one of them being an ML researcher who stated that many of the available tools in this domain are too ambitious in what they try to achieve. This leads to challenges faced during the MLOps architecture selection face, which could potentially be solved with proper predefined requirements, but it could also simply be a case of immaturity.

> *"... software development had decades and decades of maturity, and there have been lots of tools, and over time people sort of standardize on tools, but there's still big changes in software development all the time, like going to Cloud. DevOps is a good example ... And machine learning hasn't really hit that yet, so there's lots of tools and ideas on doing things a certain way, but it hasn't really been treated like typical software development. A lot of tools try to do everything in a very opinionated way. You know, follow our tool and you'll get a model deployed to an endpoint. But that doesn't work in software development generally and it wouldn't*

work in machine learning generally for like a larger project." - ID7.1

Another interviewee discussed the challenges faced in the industry related to receiving funding for implementing MLOps processes, as a result of it being so immature and its value unknown.

5.2.1.3 Data

Several interviewees mentioned data as a critical aspect of the MLOps process. The theme encompasses a range of considerations and challenges associated with data management, including technical aspects such as ingestion sources, data quality, ownership, and privacy. Additionally, the theme highlights the importance of domain knowledge and expertise in understanding and effectively working with data and the need for access to such expertise during labeling. The theme "Data" came up in 5/8 of the interviews.

Importance of Data: Traditional software projects focus on the code as the primary product, while ML projects are highly dependent on the data used to train a model. The code is merely a small part of the overall process that produces the desired outcome.

> *"Let's phrase it like this, in traditional software projects the product is really the code, but in an ML project the code is just a minor part of what actually produces the outcome because it's so dependent on the data that you used to train your models with. So two identical pieces of code used for training and inference may produce completely different results depending on the type of data you trained with..."* - ID4

Another interviewee stressed the fact that the emphasis in ML should be placed more on the data rather than the model.

> *"It is important with modeling but you can find your favorite model and you can tweak it quite good if you're really eager to do it. I mean for natural language processing or other applications there exist a lot of different models and sometimes the output you get is completely similar, but in some cases, some models perform a lot better. The same with image processing or object recognition or other things, but it's really preprocessing of images, preprocessing of measured signals, preprocessing and storing of data. Do it efficiently as well, aggregating data, etcetera. That is really the key, I would say if you're successful in that, you will be successful in ML."* - ID8

Dataset Size: The importance of data availability was discussed by interviewees. When the available data is insufficient to produce a reliable model it can be challenging to generate meaningful insights or predictions.

> *"Usually they say the machine learning gives better results than the pre-*

> *vious solution. So usually in the projects I worked on the machine learning model was better. But it happened at some time that the results that were coming from the machine learning model were on par or worse. It happened once and it's because there wasn't enough data to really build anything concrete with machine learning. So it was difficult to do something.* - ID1.1

Data Quality: Having the correct data is crucial for the success of ML projects. It is essential to carefully consider the availability and quality of data before starting development on an ML solution, as these factors can significantly impact the outcome.

> *"It's important to have the right data in most of the machine learning projects. And it might be hard to get the right data since it might be private data and then not available and so on. Maybe you don't have the right possibilities to create a good model because you don't have the right data."* - ID6.1

Data Privacy and Ownership: Considerations around privacy and data ownership should be taken into account, especially when working with different entities. While it may be less of an issue when working within the same company or group, caution must still be exercised to ensure the proper handling of sensitive data.

> *"I don't think you're talking about privacy and data ownership, but maybe it should be in the scoping. Because if you work within the same company, maybe it's not much of a problem, but yeah, if it's two different entities then it becomes quite a big problem. So like even in between the same group, you still need to be careful with how you deal with the data."* - ID1.1

Ingestion Sources: It is important to have a thorough understanding of the source of data and its corresponding metadata to assess the potential presence of biases in the data. The metadata collected should describe the data source and the data collection process. Additionally, it is essential to consider not only the origin but also the method of data collection, including whether it was an automated process or a human-based approach. By carefully analyzing both the source and collection methods of data, we can ensure its reliability and accuracy for any intended use.

> *"...So basically there should be a good understanding of the source of data and the metadata has to sort of really describe that source and that's important to see if there's any type of bias in that data. So it's not only where did it come from, but how is it collected? Like was it some automated tool that ran every Tuesday night or was it someone asking someone on the phone?"* - ID7.1

Another interviewee emphasized the importance of having a consistent ingestion

source that provides data according to what the model is trained on. Otherwise, the model may not be able to comprehend its meaning, causing challenges within the data pipeline. Additionally, attempting to skip over such data and translate it later can be a difficult task, requiring additional resources and potentially resulting in errors.

> *"Yeah, because if you train your AI to read sentences, it's not gonna understand if suddenly the sentence is only part numbers. Although it means something to the person who typed it, it doesn't mean anything to the AI, and the data pipeline. It's also hard to skip over reading that data and suddenly say, oh, this is a part number I need to translate it"* - ID1.1

Data Ingestion Cycle: The data collection approach varies depending on the domain and the specific data being gathered. In cases where time series data or data in windows of time are being collected, there may be instances where the selected window does not capture a natural cycle, leading to incomplete or inaccurate results. As such, it is essential to consider the unique characteristics of the data being analyzed and select appropriate techniques that account for any potential biases or limitations.

> *"That's one approach. It depends on the domain as well, cause sometimes there could be if you're collecting, data, time series data, or data in windows of time. Your windows of time may not capture a natural cycle, right?"* - ID7.1

Labeling: When working with supervised learning, proper labeling of data is critical, as it serves as the foundation for the model's understanding of the task at hand. Establishing clear labeling standards and guidelines is essential, with input required from both the data scientists, for technical considerations, and the labeling experts, who are typically domain experts. Thus, Ensuring accurate labeling through effective communication between these groups can help to prevent errors and biases and ultimately improve the overall quality of the data used to train the model.

> *"And then if this is supervised, the labeling things super important and the labeling standard and guidelines are really important and that should be asked to both the data scientists for technical distinctions, but more toward the labeling person, which is probably a domain expert, not related to the problem."* - ID7.1

Access to Domain Expert: Several interviewees stated that access to domain experts is vital in ensuring accurate and effective data analysis. While data analysts may possess a versatile skill set, they may not always have the necessary domain expertise to fully understand the data being analyzed. As such, it is essential to have access to subject matter experts who can provide insights and context to inform data analysis and hypothesis development. By leveraging the expertise of others, data

analysts can improve the quality of their work and make more informed decisions based on the data.

> *"Access to expertise. That's also an important thing. As a data analyst, you don't always know. So it's a very versatile role. I mean you can do data analysis in many different fields, but you don't always have the expertise in the field and you need to understand the data you're reading. So it's important to have access to experts. To understand the data and explore the data and have the right hypothesis before you start working."* - ID1.1

An interviewee highlighted the need for input from both technical professionals and domain experts while developing guidelines for data labeling. While domain experts can provide specific knowledge and context for the task at hand, technical professionals may also need to be involved to ensure that the guidelines are technically feasible and clear. By collaborating effectively, teams can create guidelines that are accurate, comprehensive, and usable, ultimately improving the quality of the data used.

> *"So for instance, if it was to recognize, is this cell cancerous, you're not gonna ask the data engineer that. It's gonna be a doctor and they're gonna have to have some sort of, like, I need a picture of it, you know, and I can circle something in the picture. So then the guidelines will have to be really specific to that domain expert. So they understand what you mean, but it could be that the data person needs to know, Hey if you circle it, don't go outside these bounds or something. You know some technical thing that they can't deal with. So it's a mix between them. So both."* - ID7.1

5.2.1.4 Infrastructure

The theme "Infrastructure" encompasses considerations and challenges associated with the infrastructure side of MLOps. All of the 8 interviewees discussed this theme, which includes the codes: Versioning, System Reliability and Scalability, Infrastructure Requirements, and Deployment Requirements. These codes relate to the importance of developing and maintaining a robust infrastructure to support data management. This includes effective versioning and storage practices, as well as the need for automation and testing to ensure the reliability and scalability of infrastructure. Additionally, the theme includes codes related to the challenges and complexities of implementing and migrating infrastructure, particularly due to the high level of technical competence necessary in order to implement it.

Versioning: Based on the insights gathered from ML and MLOps practitioners, it is evident that having the ability to access previous versions of the code is crucial for identifying and rectifying errors or malfunctioning functionalities, leading to improved development efficiency. While version-controlling code is standard practice in traditional software development, the context of ML highlights the equivalent

significance of versioning: Data, ML models, ML code, hyperparameters, and results.

> *"... when you have a bug in normal software, you go back to a version that you have released, you set it up and you can test them and see what the bug was and then test it ... But here (ML development), to spot what the error was in a model, you would need to have so many more parts that can be set up again exactly how it was, and that might not be possible if you haven't set up the pipeline in such a way so that you know exactly what guidelines, what data you had, how it was labeled, what the preprocessing step was. And maybe even you changed how you calculated the metrics, what the test set was... And then the software itself, where the model is used. So it's much more pieces that need to be traced." - ID6.1*

> *"Using a version control system is not enough for this, like Git, for example, you'll only see what you committed and what you've pushed up. But you won't see what the results are."* - ID3

System Reliability and Scalability: The significance of incorporating reliable and scalable infrastructures as a primary objective was stressed, with emphasis on MLOps' adherence to well-known DevOps methodologies such as Continuous Integration/Continuous Deployment (CI/CD), predefined test suites, and automation. The implementation of these practices is equally crucial for MLOps. By utilizing CI/CD, developers can efficiently address any unexpected problems in a consistent and regulated manner. Additionally, employing automated procedures and predefined test scenarios minimizes the risk of unintentionally introducing new bugs when attempting to add new code or fix old ones.

> *"(With CI/CD) you can deploy faster and so if there is a new feature or new bug you can quickly get it out to the customers and not have all these manual steps in between to secure that it is good and not breaking something else and it's safe to use."* - ID6.1

> *"When you have a CI/CD pipeline, you're kind of forced to conform to the tests and so you have a harder time to screw up."* - ID4

We see that the automation of ML training also was considered important, as well as the automation of other general repetitive tasks.

> *"With machine learning ... it is better to be automated because it should be run several times and to visit appropriate results and there should be an analysis done and so on. So I think it would be great if such processes can be automated." - ID2*

> *"Trying to always make it (the development process) more automated in every step. So you don't have to do manual regression testing and so on.*

> *Maybe you need to do it in parts, but then you automate as much as you can."* - ID6.1

Deployment Requirements: Towards the final parts of the MLOps life cycle, interviewees talked about various challenges related to the deployment of the developed ML system. Regardless of whether MLOps is used in the development and operation of the final ML system, practitioners face this challenge when trying to deploy to customers' environments.

> *"You can create a good machine learning model that yields good results, but the problem is the implementation into the system. It happened that it was actually complicated to fit the machine learning model with the machine (the edge device). So even though it was yielding good results. It was extremely hard to connect it directly to ... the machine."* - ID1.1

> *"... A problem, in this case, was that it was very strict what server it could run on, ... so we were training and testing on different kinds of servers and then when we actually deployed them they would be running on different servers, so to understand the impact off running it in a different environment, maybe packaged it in a different way that was not optimal, because then we could not fully test it, how it would be in a production environment and the impact of it. And then maybe there were some challenges in which libraries could be used. That could impact the model in the end."* - ID6.1

Infrastructure Requirements: The interviewees stated that it was crucial to consider what MLOps infrastructure to set up, including the right selection of tools, hardware capabilities, and if it should be hosted on the cloud.

> *"What resources are needed in terms of computing capabilities for training? Would you need the most heavy-duty GPU? Perhaps you are constrained in the beginning that you cannot use the big cluster of several GPUs then you have constraints in your choice of model."* - ID4

Additionally, one interviewee discussed the importance of reliable MLOps infrastructures, specifically when working with real-time data.

> *"With real-time data, I think infrastructure needs to be much more reliable. Because if there is an interruption, then you could lose a lot of data over it. And your system can be damaged."* - ID1.1

Lastly, if proper consideration is not taken regarding what infrastructure is necessary, then challenges regarding infrastructure migration quickly become highly relevant and costly for MLOps practitioners.

> *"I would also rather use it (the developed artifact) as a guideline to select*

my architecture or select my setup. Really choosing the right database or how you should store the data. That is a very explicit decision. It's very costly to change that." - ID8

5.2.1.5 System Monitoring

The "System Monitoring theme" was discussed by 6/8 interviewees and covered the importance of monitoring a deployed system to maintain and improve the performance of an ML system within an MLOps process. This theme includes four codes: Monitoring, Model Retraining, Experimentation Logs, and Error Analysis. These codes primarily discuss the importance of monitoring ML systems, retraining models to maintain accuracy, tracking experiments to improve reproducibility, and analyzing errors to enhance system reliability.

Monitoring: Nearly all the interviewees emphasized the significance of monitoring a deployed system as it enables a team to gain insights into the system's performance over time and improves their ability to optimize the predictive capabilities of the model.

"First of all, we created a metric system dashboard. Where we could monitor the performance of our function in real-time. And not going into too many specifics, but it was very easy to classify whether you did a correct prediction or a wrong prediction. And that was also running real-time with the software where we could actually see how we improved or deteriorated over time." - ID8

Error Analysis: Monitoring and understanding models can be a challenging task, especially when dealing with a large number of cases and multiple classes. While monitoring can help identify issues and track performance, it is equally important to understand why a model is making certain predictions, especially when they deviate significantly from the training dataset.

"So monitoring because it's so many cases and yeah, it's hard to read 1000 every week. And understanding the model because in our case we have maybe 100 classes. And sometimes the prediction is very far off, and it's hard to understand why it is so off compared to the training data set. That is sometimes challenging." - ID1.1

"...So it's not only monitoring them but also understanding why the model is giving this prediction over another one." - ID1.1

Model Retraining: Monitoring the input and output distributions of a model over time can be useful for determining when to retrain and what areas to focus on during retraining. By measuring the distribution of the input data and comparing it to previous distributions, it is possible to detect changes that may require retraining. Similarly, measuring the output distribution can help identify changes in the model's performance that may also indicate the need for retraining.

> *"There are a lot of statistical techniques on determining when say if you have a classifier for example. If you can measure the distribution of the input data and compare it over time and see if it changes. By simple metrics like mean and standard deviation, stuff like this, simple statistical metrics and you can also measure the same for the output and one way to determine if you need to retrain it is if the input is changing over time."* - ID7.1

> *"...So I would like to monitor how classification compared to the AI's current categorization and see if it matches and If there is a strong difference, I would like to be able to focus maybe the learning of the model on those differences."* - ID7.1

Experimentation Logs: Interviewees stated that when retaining a model, it is important to track different parameters or configurations in order to compare the results of the model. Thus, enabling the possibility to identify which parameter values or settings produce the greatest results. Improving the reproducibility of a model can aid in troubleshooting and diagnosing issues that arise during development or deployment.

> *"I think if you have like one thing that's necessary for machine learning models is to see the results of each. The result of this model when you use this parameter for instance. And then when you tweak that parameter, you get a totally different response. I think finding a way of tracking back and see when did I get this result, and when did I get this result, and be able to compare both would be something very vital."* - ID3

5.2.1.6 Results From the Pipeline-focused Interview

During the supplementary interview with interviewee ID1.1, the discussion centered on their team's current data pipelines (see Figure B.2 for the interview script), but no new insights were gained regarding MLOps issues. Instead, the interview reinforced the validity of the previous concerns and best practices raised in the general interviews presented earlier in this subsection. Nevertheless, the outcomes of this interview were factored into the refinement of the artifact described in subsection 5.2.2.

5.2.2 Solution candidates

Based on the results found during the problem investigation in cycle two (see subsection 5.2.1) and the evaluation feedback during cycle one (see subsection 5.1.3), the second version of the artifact was created. This artifact is found in Appendix A in Figure A.3 and Figure A.4. All changes done between the first artifact and the second artifact can be seen in Table 5.2. Additionally, an introduction page was added to the artifact providing suggestions on how to use it and information on its contents, see Figure A.2 in Appendix A.

Table 5.2: Traceability matrix presenting the changes made to the artifact based on the first cycle evaluation and second cycle problem investigation and which interviewee contributed to a specific change.

Requirements Question	Change	Interviewee ID
Scoping:		
What are the business problems and can they be solved with AI?	Deleted	ID4
What are the business problems?	Added	ID4
Can the business problems be solved with ML, how?	Added	ID4
What are the metric for success?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
What are the resources needed?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Who is the end user?	Added	ID1.1, ID6.1
How will the users interact with the model, what interface will they need?	Added	ID1.1, ID6.1
Who is the domain expert and can we access them?	Added	ID6.1, ID7.1
Data:		
Where does the data come from?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
What is the data labeling standard?	Roles changed	ID6.1, ID7.1
What meta-data should be collected?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Are there any privacy concerns regarding the data?	Added	ID1.1, ID6.1
Are there any necessary data ownership considerations?	Added	ID1.1, ID6.1, ID8
How much data is expected to be stored?	Added	ID8
When does the data become irrelevant?	Added	ID8
Are there any cyclic behaviours to the data?	Added	ID8
What is the minimum amount of data that is necessary to train the model?	Added	ID8
How will the data be acquired?	Added	ID7.1
Modeling:		
What is the model baseline?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Is it necessary to audit the model? Who should audit the model? What is the audit focus?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Which potential risks for bias exists?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
How is the input data served to the model?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
What are important business goal metrics the ML model should consider?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
What deployment constraints exist?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Deployment:		
How should the deployment process be handled?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Where should the prediction device be located?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Which software metrics are important to monitor?	Roles changed	Refined by us
Which input metrics are important to monitor?	Roles changed	Refined by us
Which output metrics are important to monitor?	Roles changed	Refined by us
How often should the model be retrained on the data gathered from deployment?	Roles changed	ID1.1, ID6.1, ID7.1, ID8
Are there any specific performance requirements?	Roles changed	ID1.1, ID6.1, ID7.1, ID8

5.2.3 Evaluation

This subsection presents the findings of the evaluation phase based on the second set of semi-structured interviews. The primary objective of these interviews was

to collect feedback on the effectiveness and usability of the revised artifact, which was developed based on the results from the previously conducted interviews and the relevant literature. The feedback collected is reported in this subsection, while the actions taken to address the feedback are presented in subsection 5.3.2. The feedback collected focuses on what the interviewees appreciated about the artifact, suggested improvements, concerns, how well the answers can translate to architecture design, how they would use it in practice, if it can lead to requirements, and how generalizable the artifact is. This information will serve as a guide for the next iteration of the artifact. Table 3.1 provides an overview of the participants involved in the evaluation stage of this cycle, where three of the interviewees were new to the study and three participated in previous interviews.

Appreciation and concerns

All of the six interviewees interviewed for this evaluation iteration expressed their appreciation for the artifact. When asked if they thought there were any benefits or drawbacks to using the artifact during an MLOps project's initial stage, all six participants agreed that it would be highly beneficial. They echoed the sentiment that the artifact could manage expectations and serve as a useful communication tool to clarify the requirements and expected outcomes of the MLOps process. Additionally, it was praised for its concrete and clear structure and content.

> *"I think it has great detail and it covers a broad range of topics really nicely."* - ID7.2

Even though all interviewees appreciated the artifact overall, there were some concerns raised. We saw that there were concerns regarding the extensiveness of the artifact, where those participants suggested that it might be unnecessary in situations where one is not aiming for the models to reach production, where the ML project simply is not large enough to make it worth implementing MLOps, or where one only would like to create a high-level overview for pitching the MLOps idea. As an example, one interviewee stated that using MLOps only for experimental ML projects could be over-engineering a simple task.

> *"One pitfall could be that you're over-engineering something that should be just a simpler experiment. When you know that you're going to serve a model and you know that there's gonna be distribution drift. Yeah, then you really need to set up this infrastructure. But if you're just gonna do simple experiments trying out different things, maybe a full-fledged like MLOps development environment. Then maybe there's a risk that you're over-engineering, but it could be still good to have this framework and so I'm not sure, but that's potentially the drawback."* - ID10

Suggestions

During the interviews, several suggestions were put forth regarding potential im-

provements to the artifact's functionality. One such suggestion was to incorporate dependencies between the Requirements Questions, allowing for a better indication of the downstream impact of changes made to one Requirement Question Answer. Another suggestion was to enable the filtering of Requirements Questions based on specific roles, which would enable each role to identify all relevant questions for them across all four stages of the artifact. In addition, a suggestion was made to implement scalable input boxes to enable the documentation of more detailed requirements. Furthermore, one interviewee suggested adding functionality for adding and removing questions in order for organizations to be able to personalize the form. Finally, it was suggested that information be included in the artifact regarding the potential impact on a system if any of the questions were left unanswered.

In addition to functionality, various suggestions regarding the data stage were proposed. One interviewee recommended adding questions concerning data leakage, which is when test data leaks into the training data of a model, ultimately leading to the overestimation of a model's performance. Furthermore, another interviewee suggested including a question about the frequency of streaming data to the model, as it could either be done for every data point, which is resource heavy, or aggregated and sent in batches. This same interviewee also suggested asking about the location of preprocessing, as it may be more suitable to perform it on the edge device rather than in the data pipeline. Two interviewees suggested asking about who should be responsible for labeling the data, whether it should be done internally or externally, such as through crowdsourcing or other means.

Answers can translate to an architecture design

During the interview, the participants were questioned about whether a fully documented version of the artifact would offer enough information to enable a responsible person to select the appropriate tools and technologies for the MLOps architecture. The response from all the interviewees was affirmative, but they noted that the quality of the answers provided in the artifact would ultimately determine the efficacy of selecting the tools and technologies for the design.

Using the artifact

When posed with the question of whether they would use the artifact or not 6/6 interviewees said that they would. Subsequently, when asked about their intended usage, one interviewee stated they would employ it as a checklist for important tasks in the initial stages, and in later stages, it would be used as an onboarding tool for new team members.

> *"Yes, especially in the beginning I would use it like a checklist just to make sure we didn't overlook anything. And then later on it can be used for documentation and especially when new people join the projects, which happens all the time. They can look at everything and get a good overview as well. And then later on it doesn't hurt to refresh your mem-*

ory. Yeah, I think it's just overall good to have documentation." - ID11

The sentiment of using the artifact as a checklist in the beginning stages was echoed by three other interviewees. Additionally, two interviewees suggested using it as a planning tool to identify crucial project roles and more accurately estimate project expenses.

> *"I think for the stuff, I've been doing with [Company]. I definitely would use it when it comes to the initial product roadmap stage. So when a product team says we want these features and at one point they think, hey, we need to use machine learning, and let's do this in machine learning. At that point, those folks generally don't have this sort of scope of work involved, so they think, OK, I'll just get a researcher, build me a model, we'll deploy it. So by doing that upfront, it really helps some with the planning and costing of stuff, they understand the teams involved, and who they have to work with. I think that getting that all cleared up upfront is always good."* - ID7.2

Lead to requirements

Another question the interviewees were asked during the evaluation was whether or not they thought the artifact, and specifically, the answers documented while using it could lead to requirements for MLOps. This question was asked to all of the participants and their answers were unanimous, 6/6 stated that they believed the answers could lead to requirements. Half of the group specifically stated that it all depended on how elaborated the answers recorded were, as answers that are not detailed enough would probably not be suitable as requirements:

> *"... Maybe some (team members) say "OK why do we need to answer these questions?". So they just answer something (vague), then it cannot be used. But if you are all aligned "we should have answers to all of this". This is good!". Then answer in such a way that it is in the right level of detail. Then it can be used."* - ID6.2

The form's Requirement Questions were noted for their straightforward and concise nature. Nonetheless, several interviewees recommended adopting an iterative approach when working with the artifact to expand upon more complex Requirement Questions:

> *"OK, so a lot of them (artifact questions) are asking for like a concrete answer. Maybe you can iterate over it, so you do a workshop first version. You see what it is and then for the next workshop - you think about it, you try to answer the questions that were not giving concrete answers, and so on."* - ID1.2

Generalizability

Since the majority of interviewees in the previous cycle were from Polestar, there was a risk that the artifact may not be applicable outside of Polestar and the automotive industry. As a result, a question was posed of whether the interviewees believed the artifact was generalizable. 6/6 interviewees, including three who were not from Polestar, agreed that it could be applied beyond Polestar and the automotive sector. One interviewee outside of Polestar stated this about the artifact's generalizability:

> *"Yeah definitely. As I say, I didn't at all think about Polestar. I kind of immediately thought about my own project currently, the previous projects that have been working with, and some other projects that I have, it is definitely applicable. It goes beyond yeah, I didn't even see any automotive here. So it's quite definitely gonna be useful outside as well."* - ID9

5.3 Cycle III Findings

This section presents the results obtained during the third and final DSR cycle. In continuity, this section includes the same three subsections as the previous cycle finding chapters: Problem investigation, Solution candidates, and Evaluation.

This cycle mainly focused on evaluating and discussing the artifact, for reasons previously elaborated on in section 3.3. Thus, the following Problem investigation subsection is slimmer compared to those in subsection 5.1.1 and subsection 5.2.1. Furthermore, this also translates into this cycle's Solution candidates and Evaluation subsection being more substantial compared to those found in previous cycles.

5.3.1 Problem investigation

During the problem investigation in cycle 2 (see subsection 5.2.1), we saw numerous themes the interviewees focused on. In this cycle's problem investigation, during the first part of the second set of semi-structured interviews, the importance of some of these themes became even more evident as they were pressed again by multiple interviewees. The themes that were found to be confirmed in this DSR cycle are presented in subsubsection 5.3.1.1. Additionally, the newly found themes are presented as individual subsubsections below.

5.3.1.1 Confirmed themes

"Project Scoping" is a theme that was discussed by the interviewees in the previous problem investigation. Multiple of the interviewees from the second set of interviews confirmed this theme by discussing many similar or identical codes such as **Understanding the Business Problem** and **Business Goals**. In addition to what was reinforced by the interviewees regarding this theme, a new code emerged, the importance of **Understanding ML**. At first glance, it may seem like an obvious requirement when working with ML. However, given that MLOps requires collaboration from a multidisciplinary team, ensuring that everyone has a comprehensive

understanding of how ML works can evidently be a challenge. This is particularly relevant to the theme "Project scoping" as multiple interviewees confirmed the importance of **Understanding the Business Problem** and **Business Goals**, which links well with the requirement to understand the underlying ML models and development processes.

"System Monitoring" is another theme that was found in the previous cycle, see subsubsection 5.2.1.5. This cycle's problem investigation found no additional codes to act as supplements to this theme. However, the data did further strengthen the validity of the theme's importance as the same topics were once again pressed.

5.3.1.2 Developing a Model

The theme "Developing a model" consists of challenges and best practices linked to model development. 2/6 interviewees discussed this theme, delving into aspects such as Data leakage, Versioning, Development environments, and CI/CD. In the previous problem investigation cycle (5.2.1), Versioning and CI/CD were covered under the "Infrastructure" theme, and the current findings did not offer any novel perspectives on these codes. Nonetheless, they validated the conclusions reached in the previous cycle. Therefore, the focus of this subsection is to present the fresh findings pertaining to the codes of data leakage and development environments.

Data leakage: One interviewee stressed the importance of avoiding data leakage while developing ML models, particularly within the research domain. Data leakage refers to a situation where information from the test set, which is meant to be used to evaluate the performance of a model, leaks into the training set, which is used to train the model. This can lead to overfitting and the model appearing to perform better than it actually would on new, unseen data. However, it was considered important outside of the context of research as well, since overestimating a model's performance can lead to organizations deploying subpar models.

> *"Yes, I think since I'm a researcher and I want to use my results to present and write papers. For me, it's very important that you don't leak test data to your model. So for me, one thing I try to be really rigorous about is data leakage. So you don't overestimate the performance of your models. "* - ID10

Development environments: Another interviewee emphasized the significance of having a development environment that closely resembles the production environment when creating a model for production use. This is essential because the performance of an ML model can be influenced by various factors, including software dependencies, hardware specifications, and operating systems. By ensuring that the development and training environment closely resembles the production environment, any difficulties or complications that may arise during the development process can be identified and addressed proactively, guaranteeing that the model functions as intended and delivers precise results.

> *"So as soon as you have a model that you want to deploy in production, basically everything from the development of the model you want to have a system or support to be able to develop it in a system which is as close as possible to the production environment. So you need to have a similar environment when you develop and train your models."* - ID9

5.3.1.3 Requirement Management

3/6 interviewees talked about the theme "Requirement Management" which consists of the codes: **Non-functional requirements**, **Dynamic environment**, and **Work continuously with requirements**. This theme discusses some important non-functional requirements to have in mind while developing production-ready ML models, the difference between pure ML projects and ML projects with MLOps introduced, and the importance of continuously working with requirements.

5.3.2 Solution candidates

Based on the results found during the problem investigation in this cycle and the evaluation feedback during cycle two (see subsection 5.2.3), the third and final version of the artifact was created. This artifact is found in Chapter 4. All changes between the second artifact and the third artifact can be seen in Table 5.3. In addition to this, an image displaying an overview of the MLOps stages and their continuous iterative nature was added to the front page of the artifact which can be seen in Figure A.5.

Table 5.3: Traceability matrix presenting the changes made to the artifact based on the second cycle evaluation and third cycle problem investigation and which interviewee contributed to a specific change.

Requirements Question	Change	Interviewee ID
Scoping:		
Is there any budget limit for the computation necessary to train the model?	Added	ID10
Data:		
Who will label the data?	Added	ID6.2, ID10
For streaming data, what is the minimum frequency of data necessary to meet business goals?	Added	ID1.2
Modeling:		
Deployment:		

5.3.3 Evaluation

The following subsection will present the findings of the final evaluation based on the workshop. This subsection includes the results in the form of feedback collected during the workshop, while the possible future actions necessary to address this feedback are discussed in Chapter 6. Furthermore, how the workshop was arranged, how it was performed, and its goal is presented in section 3.3. Lastly, the description of the imaginary case given to the workshop participants can be found in Appendix B.

The workshop was deemed a success in the sense that all of the participants could successfully discuss the scoping of the imaginary project case using the artifact, see their answers in Appendix D. According to the questionnaire answered after the workshop, the participants agreed unanimously that the artifact helped them discuss all of the necessary parts for the imaginary case. Furthermore, the participants stated that the artifact helped guide their discussion through all of the ML stages, avoiding overlooking factors they might have missed without the artifact. Thus, showcasing the artifact's effectiveness in resembling a checklist and streamlining the necessary thought process for MLOps processes. A similar sentiment was observed during the active discussions in the workshop.

Regarding the time necessary to use the artifact, it is hard to provide any definite insights. During the two-hour-long workshop, the participants managed to discuss and fill out the full artifact in relative detail. Thus, making it evident that in a situation involving a more complex case, it would be necessary to set aside more than two hours for working with the artifact.

Participants suggested that it would be beneficial to inform the artifact users that it is not necessary to answer the Requirements Questions sequentially. Although the MLOps stages are sequential, there are questions in the Scoping stage that were hard for the participants to answer before discussing and answering Requirements Questions later in other stages of the artifact, "What are the resources needed?" was one such example. This problem could potentially be improved or resolved by digitizing the artifact, which is something that was also suggested by the participants, echoing one of the suggestions from the previous evaluation iteration (see subsection 5.2.3).

Both during the workshop and in the questionnaire answers the participants recommended modifying some Requirement Questions or adding a new one to the artifact. The Requirement Question "Can the business problems be solved with ML, how?" could be extended to also ask if ML is the right approach for the said business problems. Alternatively, forming a new Requirement Question that asks this question could be another solution to this recommendation. Ultimately, it is important to discuss this question as there might exist an easier solution than developing a full ML system.

6

Discussion

The following chapter is dedicated to the discussion of the results obtained in this study. It is organized by research questions, with each section addressing one of the research questions posed at the beginning of the study. In addition, a separate section is dedicated to the discussion of the threats to validity that may have affected the reliability and generalizability of our findings. The discussions presented in this chapter aim to provide a comprehensive and critical analysis of the results, offering insights into the implications of the study and suggestions for future research.

RQ1: What are current challenges in designing an MLOps process and how do they relate to requirements knowledge?

In the first cycle problem investigation, various challenges and best practices related to requirements knowledge were discovered in the literature. The challenges identified as **P1 Data Drift** and **P2 Concept Drift** are linked to performance, which is a non-functional requirement. To address these issues, it is essential to incorporate monitoring into the MLOps architecture, which becomes a functional requirement. The significance of monitoring was emphasized by the individuals interviewed during the second and third cycle problem investigations, reinforcing its importance for operational MLOps systems as evidenced in both academic literature and expert interviews.

Similarly, **P4 Performance During Serving** also relates to the non-functional requirement of performance, with the addition of scalability. This highlights the system's need to handle the volume of traffic and maintain its performance under heavy load. This sentiment was corroborated by interviewees stressing the significance of incorporating reliable and scalable infrastructure as a primary objective, with an emphasis on MLOps' adherence to common DevOps methodologies.

The challenge **P5 Disorganized Data** discusses the difficulty of using raw data as input for training a model, especially when it is collected from different sources. This relates to the best practice **BP3 Data Quality and Labeling** which highlights the importance of proper data quality and labeling guidelines to minimize bias. Data and labeling guidelines are connected to system requirements in the capacity that system requirements might dictate the kinds of data needed for a

model and how to label that data to match these requirements. Interviews with experts further corroborated the importance of collecting enough high-quality data from consistent ingestion sources and then engaging domain experts in developing labeling guidelines.

P6 Sustainable MLOps delves into three critical components of sustainable MLOps. These are the non-functional requirements: Explainability, fairness, and accountability. While fairness was not necessarily corroborated by the conducted interviews, we see that explainability and accountability in terms of traceability were brought up by multiple interviewees. However, we feel that there is a need for further investigation of non-functional requirements related to MLOps. Nonetheless, we simultaneously see that there are both challenges and best practices related to traceability that practitioners will have to consider by using the presented artifact in this thesis.

While non-functional requirements for ML have been researched to an extent, there is very limited or no research on these requirements for MLOps. However, this makes sense in our opinion since these requirements are targeted at ML models primarily. Instead, we see that MLOps contributes to assisting practitioners to meet these requirements by bringing tools such as system monitoring, data versioning, and model versioning. By using these tools, interviewees suggest MLOps can be leveraged to meet multiple non-function ML requirements.

RQ2: Which potential solution exists to mitigate the challenges of developing an MLOps process grounded in requirements engineering?

The MLOps Requirements Form was designed based on an extensive literature analysis and interviews with relevant industry professionals, which ensures that the Requirement Questions reflect best practices in MLOps, as well as help mitigate common challenges that arise when implementing MLOps. Furthermore, since the Requirements Questions in the form were designed to work for teams regardless of their specific project or industry, it should be considered a general solution that works regardless of the selected ML model. Thus, the artifact provides a comprehensive and flexible approach to eliciting and documenting MLOps requirements. Lastly, based on our literature exploration, there exist forms targeted at technology selections for MLOps. However, a form to help elicit requirements for MLOps was not found. Therefore, to the best of our knowledge, we consider our solution to be novel.

While it may seem like the MLOps Requirements Form must be answered in sequential order, this is not necessary. There exists no strict way of using the artifact, leaving space for it to be used in various work settings such as Waterfall or Agile. If the form is employed in an Agile environment, we believed it would be particularly beneficial to digitize the artifact, making versioning of the artifact easier. This is one of our future work recommendations, which are further elaborated on in section 6.2.

In an effort to keep the MLOps Requirements Form's size reasonable, one of the

key challenges was deciding how specific the Requirement Questions could be in order to not bloat the artifact. This is because MLOps requirements can vary depending on factors such as the ML model being used, the data sources used, the target deployment environment, and the intended use case. Therefore, the MLOps Requirements Form omits some of the more non-general Requirement Questions to allow teams to tailor their MLOps requirements to their specific needs.

Despite this limitation, the MLOps Requirements Form artifact can help solve another challenge, ensuring that MLOps processes meet the intended purpose and expectations of stakeholders and developers. By using the form, the team can identify and prioritize the important MLOps requirements and establish clear and unambiguous requirements that are in line with the project objectives. As a result, the MLOps Requirements Form can help develop MLOps processes that are better aligned with stakeholder and developer expectations, ultimately leading to more successful MLOps implementations.

It is important to note that the MLOps Requirements Form is not a silver bullet solution. While it is a promising tool to help mitigate the challenges of developing an MLOps process grounded in RE, other potential solutions may also exist. Moreover, the artifact may not be suitable for every organization's needs and processes. Nonetheless, the MLOps Requirements Form is a useful tool that can significantly aid in the process of doing RE for MLOps, which is a previously overlooked step when scoping an MLOps project.

In conclusion, the MLOps Requirements Form artifact is a promising solution to the challenge of developing an MLOps process grounded in RE. Its structured and flexible approach to eliciting and documenting MLOps requirements can aid in the implementation of MLOps processes, but it is not a one-size-fits-all solution. It is one tool in a larger toolbox that organizations can use to achieve successful MLOps implementation.

RQ3: How well does the potential solution mitigate the requirements-related problems with developing an MLOps process?

Throughout this thesis, each design cycle in our DSR approach was accompanied by an evaluation iteration of the artifact. The first cycle involved conducting a set of evaluating interviews, which made it evident that the participants widely appreciated the artifact. The positive feedback received from a large majority of the interviewees led to the decision to continue iterating and improving on the initial artifact. Furthermore, the artifact was again praised in the second set of evaluating interviews, found in subsection 5.2.3.

The conducted interviews reveal that there is a gap in the industry for a tool like our artifact, which we believe partly explains the overwhelmingly positive evaluation feedback. Furthermore, our literature analysis suggests that the need for a tool to capture MLOps requirements at early stages could also be a contributing factor to

the favorable feedback received.

In addition to the positive feedback provided, the first two evaluating iterations also disclosed what could be missing, wrong, redundant, or unclear in the first two versions of the artifact, see Appendix A for previous versions of the artifact. We saw a steep decrease in comments that can be interpreted as negative in the evaluation of the second artifact, compared to the first artifact, indicating artifact completeness and saturation in terms of the broadness of the Requirement Questions.

The second set of evaluating interviews, found in subsection 5.2.3, focused on suggestions on how to make the artifact more flexible and how to use it in practice. The overarching theme for the suggestions was to digitize it in order to gain more flexibility in terms of automatically adjustable writing cells, the possibility to add one's own more scenario-specific Requirement Questions, and the ability to make the artifact generally more dynamic and customizable. Seeing how the evaluation feedback in the second cycle focused more on use cases and usability instead of its contents, as it did in the first cycle, suggests once again that the artifact is usable and covers a broad and general range of scenarios. However, we believe that the suggestions gathered in the later stages of this thesis regarding digitizing the artifact would positively impact the artifact. Unfortunately, due to the time limit, we will not be able to implement this and will instead include it as one of the suggestions for future work in section 6.2.

Finally, a workshop was held to evaluate the artifact by applying it to a simpler case, see the case in Appendix B. Unfortunately, only 3 experts were available to participate in the workshop, which made it impossible to cover all roles encompassed by the artifact. Instead, each participant took on multiple roles to compensate for this limitation. Initially, there were concerns about not having all the roles present during the evaluation. However, by assuming multiple roles the participants got a more comprehensive understanding of the artifact's capabilities and limitations, which in turn resulted in more precise feedback.

The evaluation feedback was overwhelmingly positive, two participants even stated that it should be introduced to more organizations to be used in practice. Furthermore, all participants agreed that working with the artifact fostered a collaborative environment where participants could actively engage in discussions that help in gaining a deeper understanding of both the ML problem and its solution. However, this workshop focused on a simpler case with only three participants involved. Thus, it is important to note that the results might have been different if the case would have been more complex and had a larger number of experts participating. Nonetheless, after the workshop participants were asked to answer a questionnaire about their experience working with the artifact in a group. The sentiment was very similar to what we had seen before, it helps raise discussion on topics that otherwise might have been overlooked and it would greatly help in scoping for an MLOps system. However, it should be digitized to improve the flexibility of the artifact. This again indicates that, regarding its content, the artifact is very close to being

complete and reaching sufficient saturation.

6.1 Threats to validity

We will in this section present the threats identified in this thesis, which will be categorized as internal or external. In these categories we will discuss the validity of the work process and the conclusions derived from this thesis will be examined. Additionally, the possibility of inaccurate results will also be discussed.

Internal validity

The inclusion of our company supervisor as a data point in the interviews was deemed necessary due to the limited number of individuals with experience in MLOps within the industry. However, we acknowledged potential threats to validity and mitigated them by refraining from sharing our opinions and findings with the supervisor prior to the interview.

In order to standardize the interview process, we provided interviewees with details of the subject matter and questions in advance. The interviews were conducted by either author and were equally distributed between the two authors. Furthermore, a pilot interview was conducted and then reviewed together to improve the interview protocol. All interviews were held in English, with the option of Swedish to remove any potential language barriers. The interviews were automatically transcribed and any errors were afterward corrected by the authors with the aid of recordings.

Bias in data analysis is always a threat in qualitative studies. As a result of this, our approach to qualitative data analysis involved a two-step coding approach, which entailed individual coding followed by a collaborative iteration involving in-depth discussions on code interpretation and scope. This method aimed to minimize the potential biases in the coding process by having both of us examine the data twice, with one of the iterations being a joint effort.

Lastly, another possible threat to our results is the niceness of the participants in this study, as they could have potentially held back on negative feedback and comments revolving around the artifact. Since a majority of the participants in our study work at Polestar, which is our collaborating company, there is a risk that the results were biased toward niceness. As a means of mitigating this, extra precautions were taken when designing the interview questions in order for them formulated as neutrally as possible. Therefore, we believe that this precaution, paired with clarifying the participant's anonymity at the start of each interview, minimizes this risk of affecting our results significantly.

External validity

Given the limited number of experts in the MLOps field, we opted not to utilize random sampling and instead relied on purposive and snowball sampling. However,

since the interviewees had varying domain focuses and roles, we believe the sample is representative enough.

The majority of interviewees were affiliated with a single company, which could potentially undermine the generalizability of our findings. However, we took measures to address this issue by initially conducting a literature review to inform our problem investigation. Additionally, the information gathered through the interviews aligned with the literature, suggesting our findings lean towards being generalizable. To further assess the generalizability of the artifact, six interviewees were asked to evaluate the artifact's generalization during the second interview cycle. Of these, three were outside of Polestar. As presented in subsection 5.2.3, all six interviewees agreed that the artifact was highly generalizable. This result is in line with our initial expectations, given our deliberate decision to avoid adding overly specific Requirement Questions that would only apply to certain types of ML models. Instead, we opted for more general MLOps questions that would be relevant to a broader range of projects. Lastly, during the final interviews, many of the answers were similar to previous interviews, indicating saturation in the answers.

6.2 Future work

While the artifact is currently in good condition, there are still opportunities for future improvement. To enhance the functionality of the artifact, we recommend digitizing it. This digitization would allow for additional features such as filtering, dependencies, scalable input boxes, and information detailing the potential impact on a system if the artifact's questions are left unanswered. For instance, filtering the artifact based on a user's role could help identify all relevant questions for that specific role at every stage of MLOps. Dependencies between questions could also help pinpoint downstream requirements that may be affected by any changes to the initial requirements. Additionally, scalable input boxes would allow users to write more detailed requirements within the same document rather than documenting them separately. Furthermore, including information about the impact of unaddressed requirements would help prioritize which parts of the artifact require immediate attention.

In addition, it is worth noting that there may be challenges and best practices that we have not yet considered, and further research in these areas may be beneficial. Moreover, future work could involve conducting a case study to evaluate the effectiveness of the artifact in practical scenarios where it is utilized to build MLOps architectures. This would provide valuable insights into the usefulness of the artifact in real-world settings and help identify any limitations or areas for improvement.

7

Conclusion

This study's aim was to advance the knowledge of how MLOps and RE interrelate. This is important since it enables MLOps practitioners to gain similar benefits as those attained by applying RE to traditional software development. Mainly, helping identify and document the goals and objectives of an ML project. This aids in aligning stakeholders' expectations by clearly understanding what the ML system is supposed to achieve. Moreover, to attain a more organized and reliable method of maintaining ML models in production, our aim was to produce an artifact that would serve as an itinerary for MLOps adoption from a RE viewpoint. We sought to identify the present methods for implementing MLOps processes using a DSR methodology, explore the best practices already in use and the difficulties involved in creating an MLOps process, and assess the efficacy of various solutions in addressing the identified challenges.

Our research reveals that integrating RE into MLOps processes is thought to be of great value for making sure that ML models are successfully implemented and operationalized. We found a variety of challenges and best practices related to implementing MLOps. We examined methods to mitigate these challenges and apply the best practices. The result of our research is the creation of the MLOps Requirements Form, an artifact that acts as a tool for practitioners to use when building MLOps processes from a RE standpoint.

The MLOps requirements form developed throughout our study provides a practical guide for practitioners to implement MLOps processes effectively and systematically. We received an overwhelmingly positive response from experts in the field, who validated the effectiveness and potential value of the MLOps Requirements Form. The findings we have presented provide valuable insights for practitioners in various domains on how to overcome the identified challenges associated with implementing ML models in production.

In conclusion, by offering a novel understanding of the overlap between MLOps and RE, our study makes a contribution to the growing body of knowledge revolving around RE for MLOps. The contribution we make impacts both academia and the industry, where academia benefits mostly from the explored overlap between MLOps and RE, and the industry mainly from the artifact itself. We have evaluated the

artifact theoretically in this study, what remains for future researchers is to apply the artifact in a case study to test its performance and implication in a real-world setting.

Bibliography

[1] A. S. Michael Chui Bryce Hall and A. Sukharevsky. "The state of ai in 2021". (2021), [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021 (visited on 02/16/2023).

[2] D. Kreuzberger, N. Kühl, and S. Hirschl, *Machine learning operations (mlops): Overview, definition, and architecture*, 2022. DOI: 10.48550/ARXIV.2205.02302. [Online]. Available: https://arxiv.org/abs/2205.02302.

[3] D. A. Tamburri, "Sustainable mlops: Trends and challenges", in *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 2020, pp. 17–23. DOI: 10.1109/SYNASC51798.2020.00015.

[4] A. B. Kolltveit and J. Li, "Operationalizing machine learning models - a systematic literature review", in *2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, 2022, pp. 1–8. DOI: 10.1145/3526073.3527584.

[5] I. Kumara, R. Arts, D. D. Nucci, W. J. V. D. Heuvel, and D. A. Tamburri, "Requirements and Reference Architecture for MLOps:Insights from Industry", Nov. 2022. DOI: 10.36227/techrxiv.21397413.v1. [Online]. Available: https://www.techrxiv.org/articles/preprint/Requirements_and_Reference_Architecture_for_MLOps_Insights_from_Industry/21397413.

[6] D. Pandey, U. Suman, and A. Ramani, "An effective requirement engineering process model for software development and requirements management", in *2010 International Conference on Advances in Recent Technologies in Communication and Computing*, 2010, pp. 287–291. DOI: 10.1109/ARTCom.2010.24.

[7] B. Nuseibeh and S. Easterbrook, "Requirements engineering: A roadmap", in *Proceedings of the Conference on the Future of Software Engineering*, 2000, pp. 35–46.

[8] A. Fernández-López, B. Fernández-Castro, and D. García-Coego, "Ml & ai application for the automotive industry", in *Machine Learning and Artificial Intelligence with Industrial Applications: From Big Data to Small Data*, D. Carou, A. Sartal, and J. P. Davim, Eds. Cham: Springer International Publishing, 2022, pp. 79–102, ISBN: 978-3-030-91006-8. DOI: 10.1007/978-3-030-

91006-8_4. [Online]. Available: https://doi.org/10.1007/978-3-030-91006-8_4.

[9] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, and G. Elger, "Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry", *Reliability Engineering System Safety*, vol. 215, p. 107 864, 2021, ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2021.107864.

[10] V. A. Butakov and P. Ioannou, "Personalized driver/vehicle lane change models for adas", *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4422–4431, 2015. DOI: 10.1109/TVT.2014.2369522.

[11] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: A survey of case studies", vol. 55, no. 6, 2022, ISSN: 0360-0300. DOI: 10.1145/3533378. [Online]. Available: https://doi.org/10.1145/3533378.

[12] N. Polyzotis, S. Roy, S. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: A survey", *ACM SIGMOD Record*, vol. 47, pp. 17–28, Dec. 2018. DOI: 10.1145/3299887.3299891.

[13] A. Vogelsang and M. Borg, "Requirements engineering for machine learning: Perspectives from data scientists", in *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*, IEEE, 2019, pp. 245–251.

[14] H. Villamizar, T. Escovedo, and M. Kalinowski, "Requirements engineering for machine learning: A systematic mapping study", in *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, IEEE, 2021, pp. 29–36.

[15] R. Subramanya, S. Sierla, and V. Vyatkin, "From devops to mlops: Overview and application to electricity market forecasting", *Applied Sciences*, vol. 12, no. 19, 2022, ISSN: 2076-3417. DOI: 10.3390/app12199851. [Online]. Available: https://www.mdpi.com/2076-3417/12/19/9851.

[16] L. Riungu-Kalliosaari, S. Mäkinen, L. E. Lwakatare, J. Tiihonen, and T. Männistö, "Devops adoption benefits and challenges in practice: A case study", in *Product-Focused Software Process Improvement*, P. Abrahamsson, A. Jedlitschka, A. Nguyen Duc, M. Felderer, S. Amasaki, and T. Mikkonen, Eds., Cham: Springer International Publishing, 2016, pp. 590–597.

[17] M. M. John, H. H. Olsson, and J. Bosch, "Towards mlops: A framework and maturity model", in *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, IEEE, 2021, pp. 1–8.

[18] T. Mboweni, T. Masombuka, and C. Dongmo, "A systematic review of machine learning devops", in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, IEEE, 2022, pp. 1–6.

[19] A. Ng, *Machine learning engineering for production (mlops) specialization*, Coursera, 2023 [Online]. [Online]. Available: https://www.coursera.org/specializations/machine-learning-engineering-for-production-mlops.

[20] Microsoft, *Machine learning operations (mlops) framework to upscale machine learning lifecycle with azure machine learning*, Microsoft Azure, blog, 2023 [Online]. [Online]. Available: https://learn.microsoft.com/en-us/azure/architecture/example-scenario/mlops/mlops-technical-paper.

[21] E. Knauss, "Constructive master's thesis work in industry: Guidelines for applying design science research", in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, 2021, pp. 110–121. DOI: 10.1109/ICSE-SEET52601.2021.00021.

[22] A. R. Hevner, "A three cycle view of design science research", *Scandinavian journal of information systems*, vol. 19, no. 2, p. 4, 2007.

[23] O. Doody and M. Noonan, "Preparing and conducting interviews to collect data", en, *Nurse Researcher*, vol. 20, no. 5, pp. 28–32, May 2013, ISSN: 1351-5578, 2047-8992. DOI: 10.7748/nr2013.05.20.5.28.e327. [Online]. Available: http://rcnpublishing.com/doi/abs/10.7748/nr2013.05.20.5.28.e327 (visited on 01/26/2023).

[24] R. Ørngreen and K. Levinsen, "Workshops as a research methodology", *Electronic Journal of e-Learning*, vol. 15, pp. 70–81, Apr. 2017.

[25] J. Saldaña, "The coding manual for qualitative researchers", *The coding manual for qualitative researchers*, pp. 1–440, 2021.

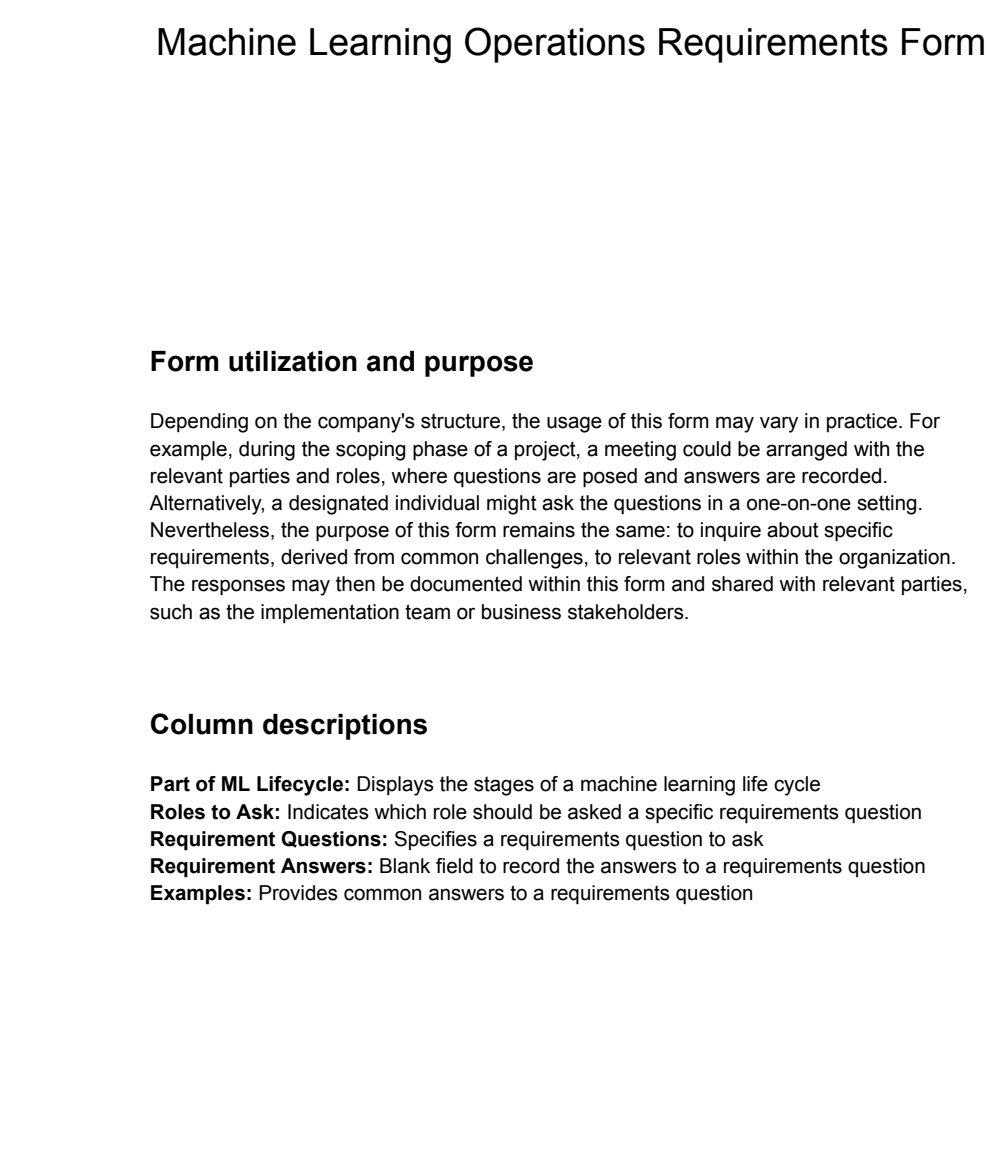
[26] L. Baier, N. Kühl, and G. Satzger, "How to cope with change?-preserving validity of predictive services over time", 2019.

A

Appendix 1

Part of ML Lifecycle	Roles to ask	Requirement Question	REQ Answers	Examples
Scoping:				
	Business stakeholder	What are the business problems and can they be solved with AI?		Has it been done before, research proves it possible, still unclear
	Business stakeholder	What are the metric for success?		ROI, customer wishes,
	Business stakeholder	What are the resources needed?		Data, time, people
Data:				
	Business stakeholder, Data scientist, Data engineer	Where does the data come from?		Owned data, crowdsourced, purchase data, purchase labels
	Data scientist, Data engineer	What data format will be used?		Structured, unstructured
	Data scientist, Data engineer	How should the data be preprocessed?		Remove data, remove duplicates
	Data scientist, Data engineer	What is the data labeling standard?		On images: Label each scratch independently on the screen, label each animal separately in the field
	Data scientist, Data engineer, Business stakeholder	What meta-data should be collected?		Time, system model, factory, device type
Modeling:				
	Business stakeholder, Data scientist	What is the model baseline?		Human-level performance, A previous system's performance, Dummy model
	Business stakeholder	Is it necessary to audit the model? Who should audit the model? What is the audit focus?		Yes/No. Business stakeholder, Third party, Data scientists. Transparency, Equality, Fairness, and Accountability..
	Business stakeholder, Data scientist, Data engineer	Which potential risks for bias exists?		Gender bias, Brand bias, Ethnicity bias
	Business stakeholder, Data scientist	How is the input data served to the model?		Batch data, Real time data
	Data scientist, IT Architect	Where should the experimental data result be stored?		Database, Excel document, JSON-file
	Business stakeholder	What are important business goal metrics the ML model should consider?		Business required classifications performance, different from general ML model performance
	Data scientist	What experimental data should be tracked?		Dataset used, Hyperparameters, Results, Results with metric summary/analysis, Training resources, Training time),
	Data scientist, Software engineer, DevOps engineer, Business stakeholder	What deployment constraints exist?		None, Edge device's hardware capabilities
Deployment:				
	Business stakeholder, MLOps engineer, DevOps engineer	How should the deployment process be handled?		Canary releases, A/B releases, Shadow releases
	Business stakeholder, MLOps engineer, DevOps engineer, Software engineer	Where should the prediction device be located?		Cloud or edge device
	DevOps engineer, MLOps engineer	Which **software metrics** are important to monitor?		Memory, computing power, latency, throughput, server load
	Data scientist, Data engineer, MLOps engineer, Software engineer	Which **input metrics** are important to monitor?		feature types (INT or String), feature range, Data schema validation
	Data scientist, Software engineer, DevOps engineer	Which **output metrics** are important to monitor?		# times users redo search, avg. prediction accuracy
	Business stakeholder, MLOps engineer, Data scientist	How often should the model be retrained on the data gathered from deployment?		Every Monday, once a month, based on deployed input/output metric triggers
	Business stakeholder, DevOps engineer, Data scientist	Are there any specific performance requirements?		Latency requirements, Query per seconds requirements

Figure A.1: First version of the artifact developed based on the literature review findings from cycle one's problem investigation.



Machine Learning Operations Requirements Form

Form utilization and purpose

Depending on the company's structure, the usage of this form may vary in practice. For example, during the scoping phase of a project, a meeting could be arranged with the relevant parties and roles, where questions are posed and answers are recorded. Alternatively, a designated individual might ask the questions in a one-on-one setting. Nevertheless, the purpose of this form remains the same: to inquire about specific requirements, derived from common challenges, to relevant roles within the organization. The responses may then be documented within this form and shared with relevant parties, such as the implementation team or business stakeholders.

Column descriptions

Part of ML Lifecycle: Displays the stages of a machine learning life cycle
Roles to Ask: Indicates which role should be asked a specific requirements question
Requirement Questions: Specifies a requirements question to ask
Requirement Answers: Blank field to record the answers to a requirements question
Examples: Provides common answers to a requirements question

Figure A.2: Front page of the second version artifact created. Supposed to serve as an introduction to the MLOps Requirements Form (the artifact).

Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Scoping:				
	Business stakeholder	What are the business problems?		Battery optimization, Fruad detection, Demand forecasting
	Data Scientist	Can the business problems be solved with ML, how?		Has it been done before, research proves it possible, still unclear
	Product owner	What are the metric for success?		ROI, customer wishes
	Product owner	What are the resources needed?		Data, time, people
	Business stakeholder	Who is the end user?		Demographical information, Internal company users, Customers
	Business stakeholder	How will the users interact with the model, what interface will they need?		App, Voice activated feature, Web page, API
	Business stakeholder, Product owner, Data scientist, Data engineer	Who is the domain expert and can we access them?		Doctors, Lawyers, Domain specific researcher
Data:				
	Business stakeholder, Product owner, Data scientist, Data engineer	Where does the data come from?		Owned data, crowdsourced, purchase data, purchase labels
	Data scientist, Data engineer	What data format will be used?		Structured, unstructured
	Data scientist, Data engineer	How should the data be preprocessed?		Remove data, remove duplicates
	Data scientist, Data engineer, Domain expert	What is the data labeling standard?		On images: Label each scratch independently on the screen, label each animal separately in the field
	Data scientist, Data engineer, Product owner	What meta-data should be collected?		Time, system model, factory, device type
	Data engineer, Legal team, Business stakeholder, Product owner	Are there any privacy concerns regarding the data?		Names, Emails, Addresses, Phone number, general GDPR concerns
	Data engineer, Legal team, Business stakeholder, Product owner	Are there any necessary data ownership considerations?		Data is owned by us, it's open source, another party owns all data
	Product owner, Data scientist, Data engineer	How much data is expected to be stored?		~10TB
	Product owner, Data scientist, Data engineer, Domain expert	When does the data become irrelevant?		Never, new product version release, annually
	Data engineer, Domain expert	Are there any cyclic behaviours to the data?		Seasonal sales cycle, full day cycle
	Data scientist	What is the minimum amount of data that is necessary to train the model?		10k images, 100 gb worth of 1080p mp3 video recordings
	Product owner, Data scientist, Data Engineer	How will the data be acquired?		Automated tool, manually collected, purchased

Figure A.3: Part one of artifact version two, see Figure A.4 for the second part. This is the form without the front page, the front page can be found in Appendix A. This artifact is the result of cycle two's problem investigation and the evaluation from cycle one.

Modeling:				
	Product owner, Data scientist	What is the model baseline?		Human-level performance, A previous system's performance, Dummy model
	Product owner, Legal	Is it necessary to audit the model? Who should audit the model? What is the audit focus?		Yes/No. Business stakeholder, Third party, Data scientists. Transparency, Equality, Fairness, and Accountability...
	Data scientist, Data engineer	Which potential risks for bias exists?		Gender bias, Brand bias, Ethnicity bias
	Product owner, Data scientist	How is the input data served to the model?		Batch data, Real time data
	Data scientist, IT Architect	Where should the experimental data result be stored?		Database, Excel document, JSON-file
	Product owner	What are important business goal metrics the ML model should consider?		Business required classifications performance, different from general ML model performance
	Data scientist	What experimental data should be tracked?		Dataset used, Hyperparameters, Results, Results with metric summary/analysis, Training resources, Training time),
	Data scientist, Software engineer, DevOps engineer, MLOps engineer	What deployment constraints exist?		None, Edge device's hardware capabilities
Deployment:				
	Product owner, MLOps engineer, DevOps engineer	How should the deployment process be handled?		Canary releases, A/B releases, Shadow releases
	Product owner, MLOps engineer, DevOps engineer	Where should the prediction device be located?		Cloud or edge device
	DevOps engineer, MLOps engineer, Software engineer	Which software metrics are important to monitor?		Memory, computing power, latency, throughput, server load
	Data scientist, Data engineer, MLOps engineer	Which input metrics are important to monitor?		feature types (INT or String), feature range, Data schema validation
	Data scientist, Software engineer, MLOps engineer	Which output metrics are important to monitor?		# times users redo search, avg. prediction accuracy
	Product owner, MLOps engineer, Data scientist	How often should the model be retrained on the data gathered from deployment?		Every Monday, once a month, based on deployed input/output metric triggers
	Product owner, DevOps engineer, Data scientist	Are there any specific performance requirements?		Latency requirements, Query per seconds requirements

Figure A.4: Part two of artifact version two, see Figure A.3 for the first part. This is the form without the front page, the front page can be found in Appendix A. This artifact is the result of cycle two's problem investigation and the evaluation from cycle one.

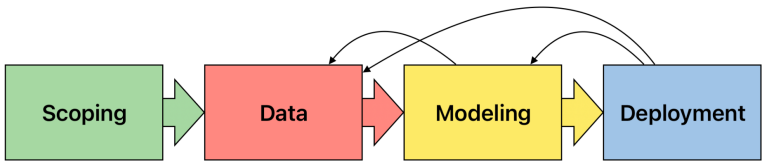
Machine Learning Operations Requirements Form

Form utilization and purpose

Depending on the company's structure, the usage of this form may vary in practice. For example, during the scoping phase of a project, a meeting could be arranged with the relevant parties and roles, where questions are posed and answers are recorded. Alternatively, a designated individual might ask the questions in a one-on-one setting. Nevertheless, the purpose of this form remains the same: to inquire about specific requirements, derived from common challenges and best practices, to relevant roles within the organization. The responses may then be documented within this form and shared with relevant parties, such as the implementation team or business stakeholders. The documented answers can be interpreted as informal MLOps requirements. Therefore, they can be used as they are or as the foundation for creating more formal requirements depending on the implementation context.

The MLOps stages

MLOps is an iterative framework that requires constant maintenance and monitoring. Therefore, it is common for the MLOps requirements to evolve and change iteratively in parallel. The figure below gives a visual representation of how the changes and information from one stage feed into another.



Column descriptions

Part of ML Lifecycle: Displays the stages of a machine learning life cycle
Roles to Ask: Indicates which role should be asked a specific requirements question
Requirement Questions: Specifies a requirements question to ask
Requirement Answers: Blank field to record the answers to a requirements question
Examples: Provides common answers to a requirements question

Figure A.5: Front page of the final artifact created. Supposed to serve as an introduction to the MLOps Requirements Form (the artifact).

B

Appendix 2

1st set of semi-structured interviews

Introduction and explanation of the study (Max 5 minutes)

Demographics and Professional Experiences (Max 10 minutes)

1. Can you shortly introduce yourself and tell me which team you work with?
2. How long have you worked in the industry?
3. What is your current role in the company?
4. What does this role entail? (What usually is your part of projects?)
5. Does your education match your current role and what is your education?

Open interview (Max 25 minutes)

General Questions:

1. Have you worked with requirements for software projects? Describe your experiences briefly.
2. Have you worked with ML projects?
    a. Was the project deemed successful?
        i. If the project was successful, how did you know that the project was successful?
        ii. If it was not successful, what indicated its failure?
3. Have you worked with DevOps on projects? Describe your experiences briefly.
4. Have you worked with or thought about using DevOps for ML projects (MLOps), what are or would be the challenges?
5. Have you worked with or thought of specifying requirements for ML projects? What were or could be the challenges?
    a. If worked with: what worked/works well?
6. Would the MLOps process need further requirements than typically available for ML projects?

Last question:

1. Is there anything you would like to add? Any factors you think we have missed, or something else you want to add to the interview?

Evaluation of prototype artifact from literature (Max 20 min)

1. The following are MLOps requirement questions elicited from the literature. Do you consider them helpful, please motivate why or why not?
2. Can you spot anything that is missing, wrong, or redundant?
3. Do you believe that you and your colleagues would benefit from using this form/guide when scoping for a new MLOps process/infrastructure?
    a. If that is the case, in which way?
4. Do you have any suggestions for improvements to the artifact? These could be related to the structure, UX, content, or anything else.

Figure B.1: The script used for the first set of semi-structured interviews.

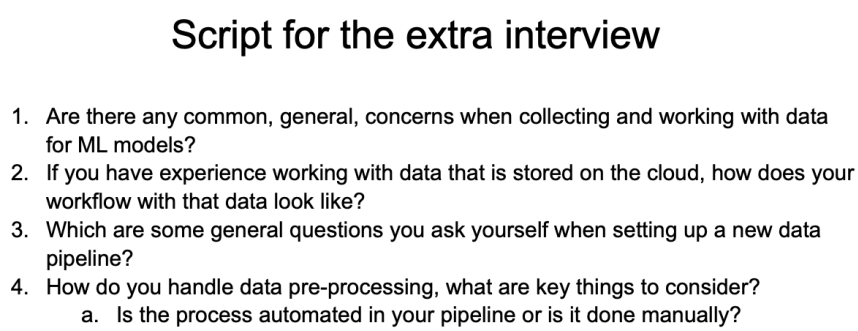


Figure B.2: The script used for the extra interview held during iteration 2 which focused on the participants' team's current data pipelines and workflows.

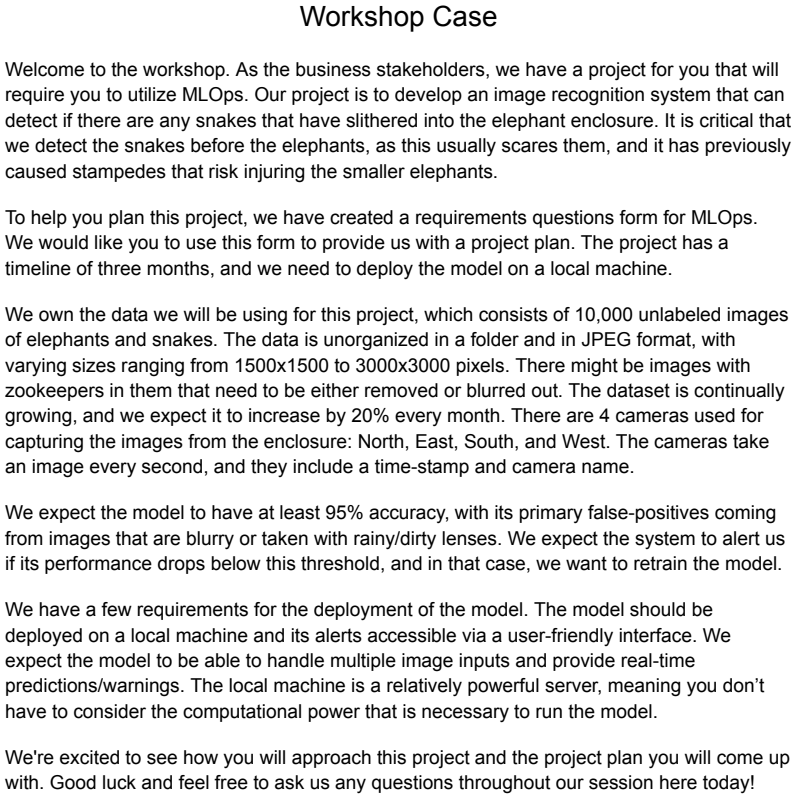


Figure B.3: Description of the imaginary case given to the participants during the workshop.

List of Requirements

Data Description:

- The data consists of 10,000 unlabeled images of elephants and snakes, unorganized in a folder.
- The project team is responsible for labeling the data.
- The data is in JPEG format with varying sizes ranging from 1500x1500 to 3000x3000 pixels.
- The Zoo owns the data.
- The dataset is continually growing, and we expect it to increase by 20% every month.
- The camera used for capturing the images takes an image every second.
- Some images might have zookeepers in them that need to be either removed or blurred out.

Model Requirements:

- The model should achieve at least 95% accuracy on the validation set, with its primary false-positives coming from images that are blurry or taken with rainy/dirty lenses.
- The model should be able to handle multiple image inputs and provide real-time predictions/warnings.

Deployment Requirements:

- The model should be deployed on a local machine.
- The model's alerts should be accessible via a user-friendly interface.
- The local machine is a relatively powerful server, meaning you don't have to consider the computational power necessary to run the model.

System Requirements:

- The system should be able to handle the growing dataset and ensure that the model remains up to date.
- The system should be able to track the performance of the model over time and provide alerts if performance drops below a certain threshold.

Figure B.4: Requirements sheet for the imaginary case given to the participants during the workshop. This was used as a summary of the case during the workshop.

Evaluating Workshop Questions

1. Were the questions in the form clear and unambiguous, or were there any areas where further clarification was needed? If so, where?

2. Were there any challenges or barriers to working with the form as a group, and if so, what were they and did you overcome them somehow?

3. Did you find the form useful or not for guiding discussions and ensuring that all relevant areas of MLOps were covered in your group's work? Is anything missing?

4. Was the form helpful or not in promoting a shared understanding of the MLOps requirements for the image recognition system among the members of your group? If so, why? if not, why not?

5. Were there any areas where you felt that additional guidance or instructions would have been helpful in working with the form with your group? If so, where?

6. Were there any areas where you felt that the form could be improved or modified to better suit your group's needs? If so, where?

7. Overall, did you or did you not find the MLOps requirements form to be a helpful and effective tool for working collaboratively with your group to identify and document MLOps requirements for an image recognition system? Why, or if not, why not?

Figure B.5: Set of evaluating questions given to the participants after the workshop.

C

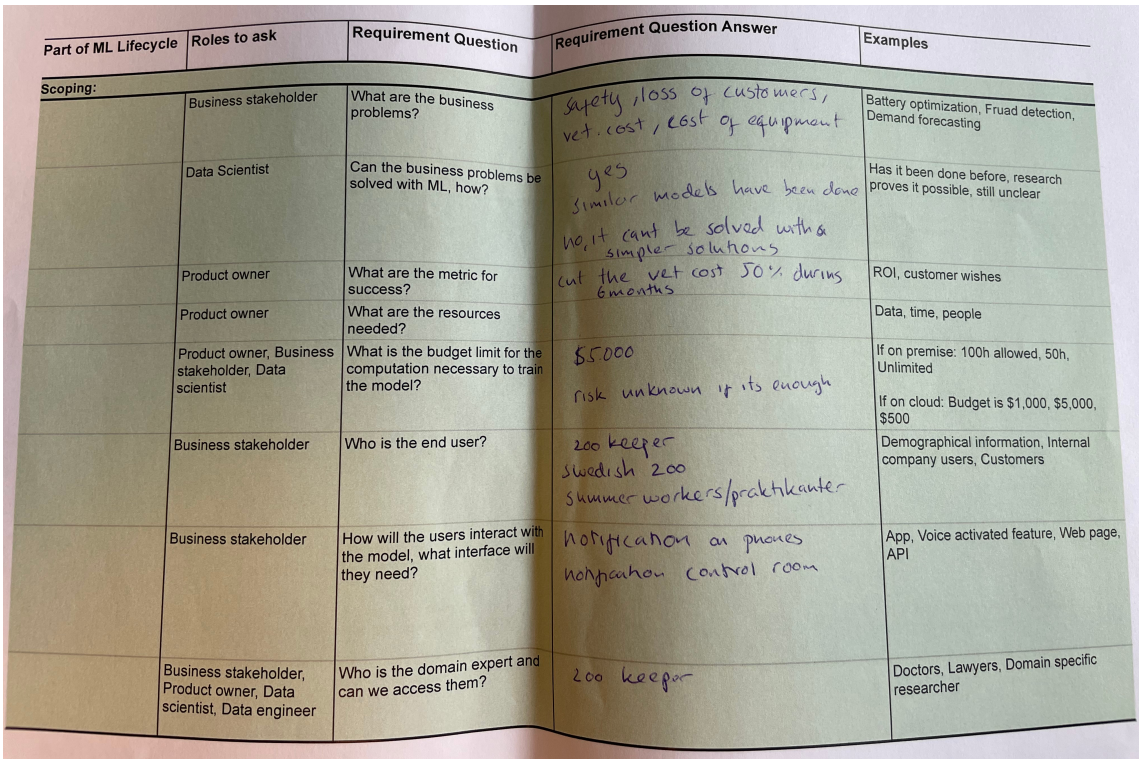
Appendix 3

Table C.1: Codebook displaying the collection of inductive and deductive codes and their description used during the thematic analysis.

Codes:	Description:
Access to domain expert	The ability of data scientists to consult with subject matter experts who have deep knowledge of the domain or industry in which the machine learning model is being applied.
Automation	The use of software tools and algorithms to automate the process of building, training, and deploying models.
Avoid development requirements	Requirements specifying how development should be done is inadvised
Business goals	The specific objectives that a company or organization is trying to achieve through the use of machine learning technology.
CI/CD	Set of best practices and tools for automating the development, testing, and deployment of models.
Customer feedback loop	Process of continuously gathering feedback from users or customers of a machine learning product or service, and using that feedback to improve the performance and usability of the product or service.
Data ingestion cycle	Collection of a complete data cycle
Data leakage	Information from the training dataset is inadvertently included in the test dataset or otherwise used to inform model development.
Data ownership	The legal and ethical ownership and control of the data used in models. This includes considerations such as who owns the data, who has the right to access and use the data, and how the data can be used.
Data privacy	The protection of sensitive and personal data used in ML models.
Data quality	The accuracy, completeness, and consistency of the data used to train and test ML models.
Dataset size	The amount of data necessary to train and test a machine learning model.
Deployment requirements	The specific needs and constraints for deploying a machine learning model in a real-world production environment.
Development environment	The set of tools, software, and hardware used to develop, test, and refine machine learning models
Difficult to evaluate if requirements are met	Difficult to pinpoint when a requirements in an ML environment is accomplished
Documenting	The process of creating and maintaining comprehensive and accurate documentation for a ML project.
Dynamic environment	Environment in which the data or conditions may change over time.
Error analysis	The process of analyzing the errors made by a model to identify patterns or trends that can be used to improve its performance.
Experimentation logs	Systematic and comprehensive record of experiments that have been conducted during the development and optimization of machine learning models.
Given requirement	Requirements that are elicited by another person and then given for implementation
Hard to implement infrastructure	Diffuclt to implement MLOps infrastructure
Importance of Data	Recognition of the impact data has on the success of an ML model
Infrastructure migration	The process of transferring ML models and associated workflows from one infrastructure environment to another.
Infrastructure requirements	Configuartion decisions regarding the infrastructure
Ingestion sources	The various types of data sources that can be used to feed data into ML models
Inter-team communications,	Communications within a single team or between multiple different teams
Labeling	The process of assigning a categorical or numerical value to a data point or sample.
Manage expectations	The process of setting realistic goals and outcomes for a ML project, communicating those goals to stakeholders, and regularly evaluating and adjusting those expectations as the project progresses.
Maturity	The level of sophistication and effectiveness of organizations ML operations processes and practices.
Model baseline	A simple, minimal or trivial model that is used as a benchmark to evaluate the performance of more complex models.
Model requirements	The specifications and expectations that a model must meet in order to be considered successful and useful for its intended purpose.
Model retraining	The process of updating or refining a model using new data or updated parameters.
Monitoring	Continously observing the performance of a model over time to ensure that it is still accurate and relevant for its intended task.
Non-functional requirements	Characteristics and qualities of a system that are not related to its primary function or task, but rather to its overall performance, scalability, reliability, and maintainability.
Productionize model	The process of deploying a machine learning model into a production environment, where it can be used to make predictions on new data in real-time.
Self-made requirement	Requirements that are elicited by the implementation team
Testing	The process of evaluating ML models to ensure that they are working as intended and producing accurate results.
Understand ML	Understanding of how machine learning works
Understand the data	Understanding the data needed to train a model
Understand the end-user	Understanding the end-users of a model
Understand the problem	Understanding the problem being solved with ML
Value of MLOps	The value MLOps brings to an organization
Versioning	The practice of tracking and managing changes to ML models and associated artifacts over time. This includes tracking changes to the model code, data sets, model configurations, and other related resources.
Work continously with requirements	A dynamic system requires continous work on requirements

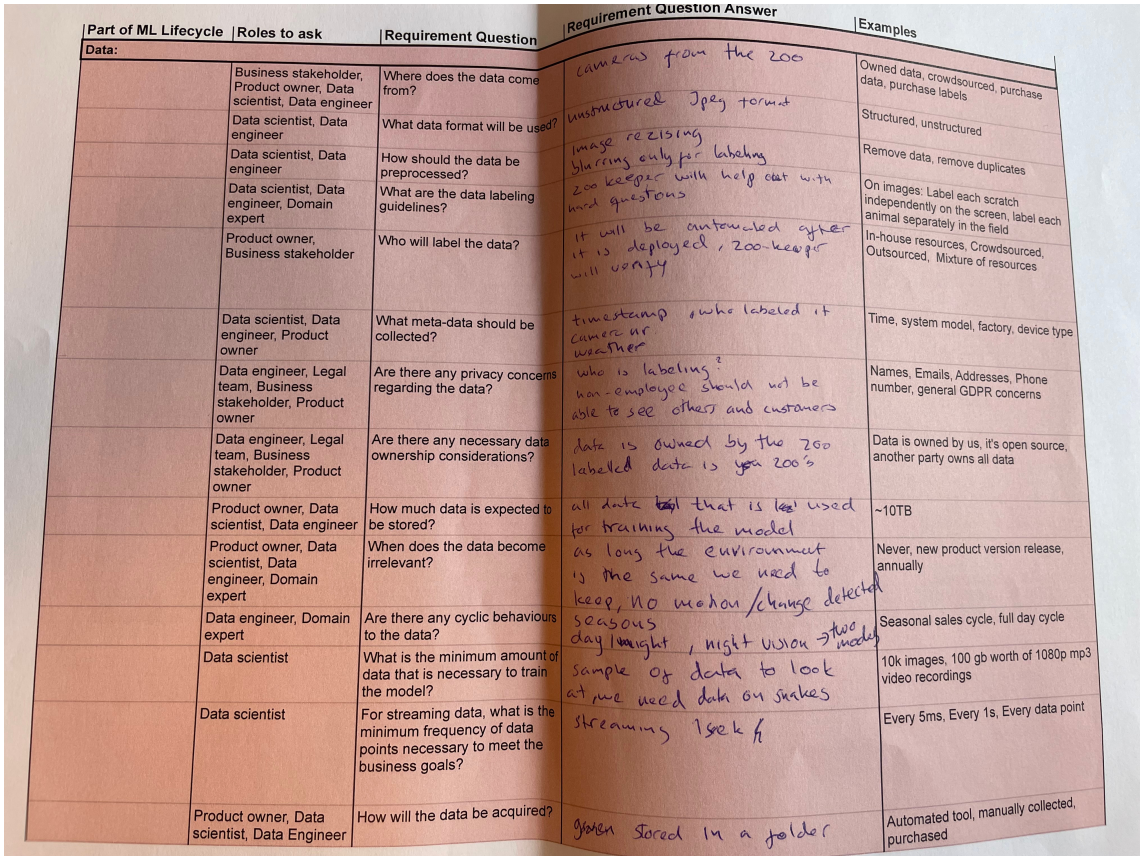
D

Appendix 4



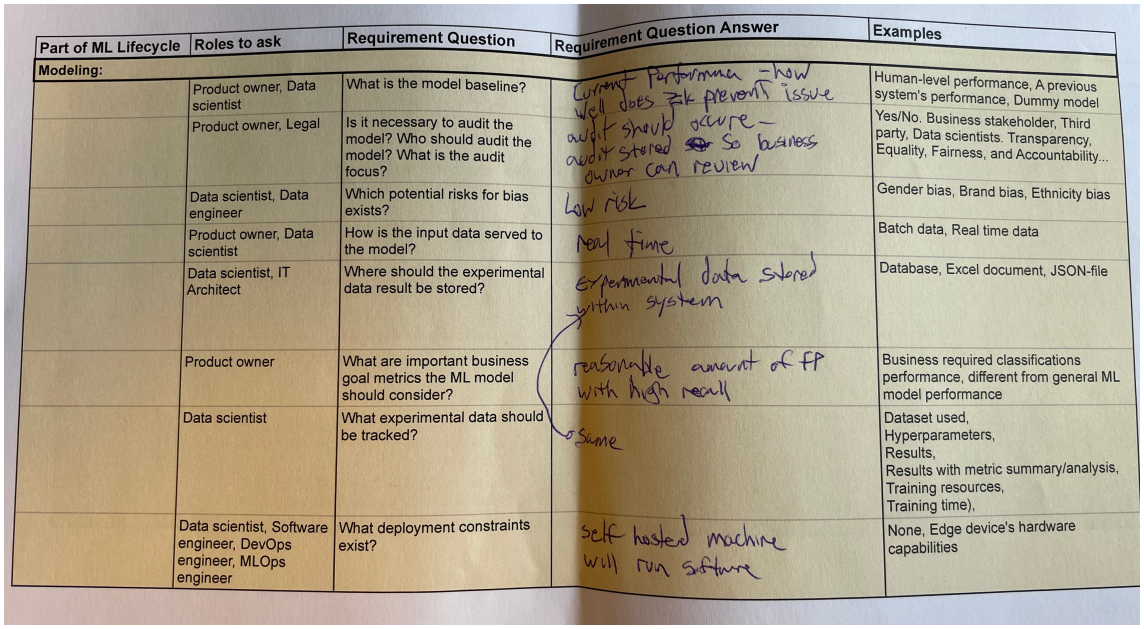
Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Scoping:				
	Business stakeholder	What are the business problems?	safety, loss of customers, vet. cost, cost of equipment	Battery optimization, Fruad detection, Demand forecasting
	Data Scientist	Can the business problems be solved with ML, how?	yes similar models have been done no, it cant be solved with a simpler solutions	Has it been done before, research proves it possible, still unclear
	Product owner	What are the metric for success?	cut the vet cost 50% during 6months	ROI, customer wishes
	Product owner	What are the resources needed?		Data, time, people
	Product owner, Business stakeholder, Data scientist	What is the budget limit for the computation necessary to train the model?	$5000 risk unknown if its enough	If on premise: 100h allowed, 50h, Unlimited If on cloud: Budget is $1,000, $5,000, $500
	Business stakeholder	Who is the end user?	zoo keeper Swedish zoo Summer workers/praktikanter	Demographical information, Internal company users, Customers
	Business stakeholder	How will the users interact with the model, what interface will they need?	notification on phones notification control room	App, Voice activated feature, Web page, API
	Business stakeholder, Product owner, Data scientist, Data engineer	Who is the domain expert and can we access them?	zoo keeper	Doctors, Lawyers, Domain specific researcher

Figure D.1: Workshop participants Requirements Questions Answers regarding the scoping stage of the imaginary case.



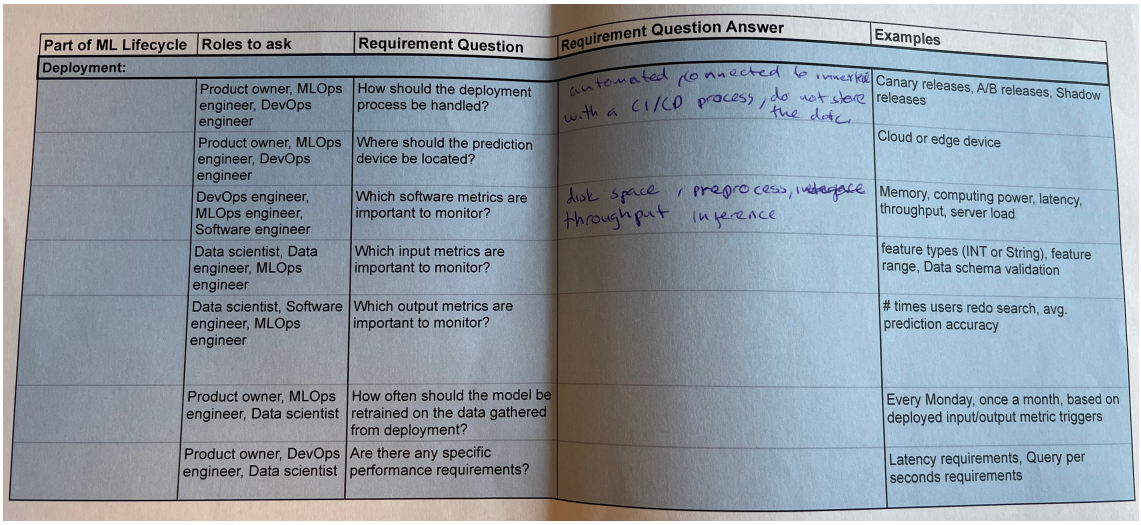
Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Data:				
	Business stakeholder, Product owner, Data scientist, Data engineer	Where does the data come from?	cameras from the zoo	Owned data, crowdsourced, purchase data, purchase labels
	Data scientist, Data engineer	What data format will be used?	unstructured jpeg format	Structured, unstructured
	Data scientist, Data engineer	How should the data be preprocessed?	image resizing blurring only for labeling zoo keeper will help out with hard questions	Remove data, remove duplicates
	Data scientist, Data engineer, Domain expert	What are the data labeling guidelines?		On images: Label each scratch independently on the screen, label each animal separately in the field
	Product owner, Business stakeholder	Who will label the data?	it will be outsourced after it is deployed, zoo-keeper will verify	In-house resources, Crowdsourced, Outsourced, Mixture of resources
	Data scientist, Data engineer, Product owner	What meta-data should be collected?	timestamp, who labeled it camera nr weather	Time, system model, factory, device type
	Data engineer, Legal team, Business stakeholder, Product owner	Are there any privacy concerns regarding the data?	who is labeling? non-employee should not be able to see others and customers	Names, Emails, Addresses, Phone number, general GDPR concerns
	Data engineer, Legal team, Business stakeholder, Product owner	Are there any necessary data ownership considerations?	data is owned by the zoo labeled data is yes zoo's	Data is owned by us, it's open source, another party owns all data
	Product owner, Data scientist, Data engineer	How much data is expected to be stored?	all data that is used for training the model	~10TB
	Product owner, Data scientist, Data engineer, Domain expert	When does the data become irrelevant?	as long the environment is the same we need to keep, no motion/change detected	Never, new product version release, annually
	Data engineer, Domain expert	Are there any cyclic behaviours to the data?	seasons day/night, night vision → two models	Seasonal sales cycle, full day cycle
	Data scientist	What is the minimum amount of data that is necessary to train the model?	sample of data to look at, we need data on snakes	10k images, 100 gb worth of 1080p mp3 video recordings
	Data scientist	For streaming data, what is the minimum frequency of data points necessary to meet the business goals?	streaming 1sek ?	Every 5ms, Every 1s, Every data point
	Product owner, Data scientist, Data Engineer	How will the data be acquired?	stored in a folder	Automated tool, manually collected, purchased

Figure D.2: Workshop participants Requirements Questions Answers regarding the data stage of the imaginary case.



Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Modeling:				
	Product owner, Data scientist	What is the model baseline?	Current performance - how well does it prevent issue	Human-level performance, A previous system's performance, Dummy model
	Product owner, Legal	Is it necessary to audit the model? Who should audit the model? What is the audit focus?	audit should occure - audit stored so business owner can review	Yes/No. Business stakeholder, Third party, Data scientists. Transparency, Equality, Fairness, and Accountability...
	Data scientist, Data engineer	Which potential risks for bias exists?	Low risk	Gender bias, Brand bias, Ethnicity bias
	Product owner, Data scientist	How is the input data served to the model?	real time	Batch data, Real time data
	Data scientist, IT Architect	Where should the experimental data result be stored?	Experimental data stored within system	Database, Excel document, JSON-file
	Product owner	What are important business goal metrics the ML model should consider?	reasonable amount of FP with high recall	Business required classifications performance, different from general ML model performance
	Data scientist	What experimental data should be tracked?	Same	Dataset used, Hyperparameters, Results, Results with metric summary/analysis, Training resources, Training time),
	Data scientist, Software engineer, DevOps engineer, MLOps engineer	What deployment constraints exist?	self hosted machine will run software	None, Edge device's hardware capabilities

Figure D.3: Workshop participants Requirements Questions Answers regarding the modeling stage of the imaginary case.



Part of ML Lifecycle	Roles to ask	Requirement Question	Requirement Question Answer	Examples
Deployment:				
	Product owner, MLOps engineer, DevOps engineer	How should the deployment process be handled?	automated connected to innerfrid with a CI/CD process, do not store the data	Canary releases, A/B releases, Shadow releases
	Product owner, MLOps engineer, DevOps engineer	Where should the prediction device be located?		Cloud or edge device
	DevOps engineer, MLOps engineer, Software engineer	Which software metrics are important to monitor?	disk space, preprocess, ~~interface~~ throughput inference	Memory, computing power, latency, throughput, server load
	Data scientist, Data engineer, MLOps engineer	Which input metrics are important to monitor?		feature types (INT or String), feature range, Data schema validation
	Data scientist, Software engineer, MLOps engineer	Which output metrics are important to monitor?		# times users redo search, avg. prediction accuracy
	Product owner, MLOps engineer, Data scientist	How often should the model be retrained on the data gathered from deployment?		Every Monday, once a month, based on deployed input/output metric triggers
	Product owner, DevOps engineer, Data scientist	Are there any specific performance requirements?		Latency requirements, Query per seconds requirements

Figure D.4: Workshop participants Requirements Questions Answers regarding the deployment stage of the imaginary case.