Journal of Data Intelligence, Vol. 2, No.1 (2021) 001-029

REVEALING CHALLENGES WITHIN THE APPLICATION OF MACHINE LEARNING SERVICES – A DELPHI STUDY

ROBERT PHILIPP

University of Vienna
robert.philipp.univie@gmail.com

ANDREAS MLADENOW

University of Vienna
andreas.mladenow@univie.ac.at

CHRISTINE STRAUSS

University of Vienna
christine.strauss@univie.ac.at

ALEXANDER VOELZ

University of Vienna
alexander.voelz@univie.ac.at

Over the past years, Machine Learning has been applied to an increasing number of problems across numerous industries. However, the steady rise in the application of Machine Learning has not come without challenges since companies often lack the expertise or infrastructure to build their own Machine Learning systems. These challenges led to the emergence of a new paradigm, called *Machine Learning as a Service*. Scientific literature has mainly analyzed this topic in the context of platform solutions that provide ready-to-use environments for companies. We recently have developed a platform-independent approach and labeled it *Machine Learning Services*. The aim of the present study is to identify and evaluate challenges and opportunities in the application of Machine Learning Services. To do so, we conducted a Delphi Study with a panel of machine learning experts. The study consisted of three rounds and was structured according to the five steps of the *Data Science Lifecycle*. A variety of challenges from the areas "Communication", "Environment", "Approach", "Data", "Retraining, Testing, Monitoring and Updating", "Model Training and Evaluation" were identified. Subsequently, the challenges revealed by the Delphi Study were compared with previous work on Machine Learning as a Service, which resulted from a structured literature review. The identified areas serve as possible future research fields and give further implications for practice. Alleviating communication issues and assessing the business IT infrastructure prior to the machine learning project are among the key findings of our study.

Keywords: Machine Learning as a Service, MLaaS, Machine Learning Services, Machine Learning, Delphi study, Data Science Lifecycle, Machine Learning Platform

1 Introduction

In his pioneering work "Computing Machinery and Intelligence", Alan Turing first introduced the term and idea of a *learning machine*, i.e., a machine, that can change its internal rules of operation based on

its input [56]. Based on this, several other important approaches and enhancements were developed, such as neural networks [9], perceptrons [43], backpropagation [25, 45], reinforcement learning [54] and many others, which have led to even more scientific breakthroughs – most recently to a deep learning algorithm solving the problem of protein folding [49]. Due to advances in computing power [3], an increase in generated and available data, and overall connectivity [19], machine learning (ML) is now applied to a wide range of problems in a variety of industries. Additionally, the ability to store large amounts of data and access them from every location with an internet connection, has been greatly enhanced by developments in cloud computing. These developments enable the creation of new, usage-based business models through which companies can increase their efficiency and cost effectiveness [31]. Machine Learning as a Service (MLaaS) is such a model, that allows leveraging the technology for companies, that lack the computational and knowledge resources to develop their own solutions [42]. Without the as a Service-model, ML is typically applied in a project setting aimed at solving a specific problem. In a typical ML project different actors and stakeholders from different departments need to collaborate in diverse project steps such as data understanding, modeling, deployment, and so on [29]. While, in traditional programming where answers are derived from *rules* and data, the paradigm is different for machine learning. Here, the rules are derived from *answers* and data. With its increased application, new obstacles can be noticed that hinder the full utilization of the technology. Collaborations between different actors from different departments and the programming paradigm change combined with emerging technologies, lead to new challenges and opportunities in machine learning projects. To address the problem, this study aims to identify these challenges and opportunities by conducting a Delphi study among machine learning experts in various roles from different industries. Based on the identified challenges, new areas of research will be proposed. Our results can further enable businesses to prepare themselves better for the application of machine learning by highlighting the challenges that need to be overcome.

The paper at hand represents an extended version of [33], where we have conducted a systematic literature review on the topic of MLaaS. Section 2 briefly recaps the most important insights from that literature review. In Section 3, the methodology of the present study is described. The results obtained through the Delphi Study are presented and interpreted according to the steps of the data science lifecycle in Section 4. A discussion of the results follows in section 5. Finally, section 6 concludes our work with ideas for future research and the limitations of our study.

2 Theoretical Background of MLaaS and Machine Learning Services

The following section elaborates on selected theoretical constructs and elements that are relevant in the context of the present study, and that support the understanding of identified challenges in that field. A more detailed elaboration can be found in a prior contribution on the topic of Machine Learning as a Service [33]. First, two machine learning lifecycles are discussed. Next, we summarize the most important findings of the contributions on MLaaS, which we identified during our previously conducted systematic literature review. As a result, we were able to group the selected literature into four key concepts, i.e., Platform, Applications, Performance Enhancements and Challenges.

2.1 Machine Learning Lifecycles

This subsection presents two lifecycle models for ML projects. The lifecycles are process models that provide standards for carrying out projects and are useful for planning, communication, and documentation [60]. We elaborate on the first widely accepted data-science-related lifecycles, i.e. CRISP-Data Mining Lifecycle, and a further extended model, i.e. Team Data Science Lifecycle.

2.1.1 CRISP Data Mining Lifecycle. In 1999 the CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining was published. It serves as both, a methodology covering descriptions of the project phases, tasks and the relationships between the tasks, and a process model providing a visualized overview of the Data Mining life cycle. The model is structured into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [60]. Data Mining is known as the process in which algorithms are applied to identify patterns and relationships in large sets of data to extract previously unknow knowledge [61]. It is a subset of data science and can utilize ML methods as well. That is why the CRISP DM represents a predecessor of the data science lifecycle.

2.1.2 Team Data Science Lifecycle. The Team Data Science Lifecycle (TDSL) is an extension of the CRISP-DM lifecycle. It has been designed by Microsoft for data science projects, which deploy ML or AI models as part of intelligent applications [29]. An overview of TDSL is shown in figure 1. Compared to the CRISP-DM, Data Understanding and Data Preparation are combined into one step and Evaluation becomes part of the Modeling step. Furthermore, it adds the Customer Acceptance step after the Deployment phase. The TDSL is of importance for this work because it can be used to visualize the difference between our definition of MLaaS and the platform setting, which is commonly employed by the majority of the MLaaS research. The platform setting of MLaaS will be discussed in subsection 2.2.1, while an adapted definition is introduced in subsection 2.2.2. In addition, the TDSL is used to structure the questionnaire as well as the results of our study.

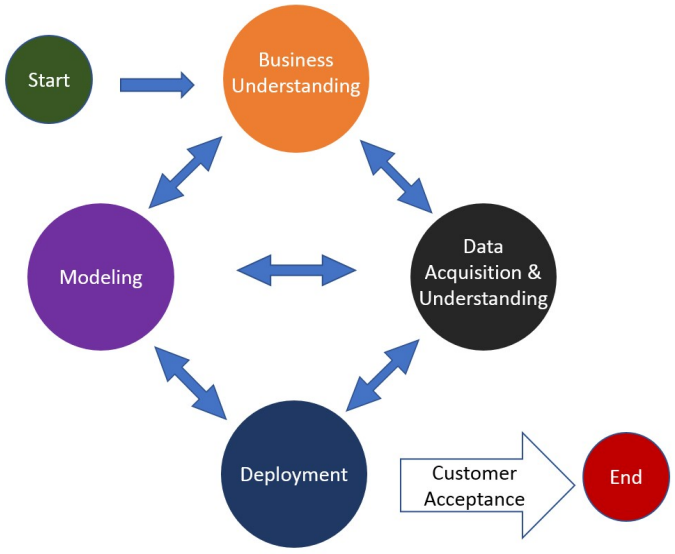

Figure 1 Team Data Science Lifecycle (adapted from [29]).

2.2 Machine Learning as Service

In our previously conducted systematic literature review [33], we identified 30 contributions on the topic of MLaaS, which we grouped into the four concepts Platform, Applications, Performance Enhancements and Challenges. The respective 30 contributions of our systematic literature research are marked with an asterisk (*) in the references. In the following subsection, we summarize the most important insights for each of the concepts.

2.2.1 Machine Learning as Service Platform. Our literature review found only limited prior work focussing on MLaaS. More than half of the 30 identified publications haven been published in 2019 or later. The term was first mentioned in scientific literature in 2016 by Ribeiro et al. [42]. The authors presented an architectural design for a MLaaS platform. Their goal was to give small companies, developers, and researchers, who lack the resources for their own ML solutions, access to a practical MLaaS platform which can be used easily. Their presented architecture focused on predictive modeling and is scalable, flexible and can support different data sources. It can further create different models with various algorithms and parameters. In the following years, several authors describe the utilization of specific MLaaS platforms that enable the sharing, versioning and validating of ML models. Zhang et al. [65] present *DataLab*, a system that manages the big data analytical workflow. It allows revision of code and data within one system and further comprises a code execution engine. In 2019 Zhao et al. [66] presented *Acumos*, a platform that enables the sharing of ML models by dividing them into microservices. Rao et al. [39] developed *Bodhisattva*, a containerized Platform as a Service (PaaS) System, which gives users the opportunity to develop, improve and deploy ML models with ease. Through their platform, they achieve faster time-to-market, model-reusability, scalability, availability, and security. Rao et al. [39] justify the need for MLaaS platforms by a lack of expertise within businesses and a gap between demand and availability of ML software solutions. An important player in this platform setting are cloud providers. Yousif [64] states that the major cloud providers are enhancing their products with functionalities that enable users to perform cognitive tasks such as ML projects. The cloud providers are obviously a good fit for MLaaS platforms since they already control the vast volumes of data. Furthermore, they already provide the vast computational power, that is needed for complex ML tasks [64]. Yao et al. [63] compared multiple MLaaS platforms to evaluate their effectiveness regarding user control and performance. The authors further analysed if higher control leads to higher quality models and identified a strong correlation between system complexity and optimized classification performance. Optimized classification performance means, that out of all possible combinations of the controls, the best performing one is chosen [63].

Overall, our research identified the need for a clearer definition for MLaaS. Within the platform setting, MLaaS-Platforms (MLaaSP) provide the ability to build, train and deploy ML models in a single toolset. This service should therefore be called MLaaSP. This definition was developed by combining our analysis of the MLaaS platforms within the scientific literature, such as Ribeiro et al. [42], Zhang et al. [65], Rao et al. [39], Zhao et al. [66], with an analysis of the capabilities of seven widely used MLaaS platforms. The selected platforms are AWS Sagemaker, Azure Machine Learning, Google AI Platform, H2O, IBM Watson Machine Learning, Oracle AI and Alibaba Machine Learning Platform. Our comparison concluded that all platforms provide the ability to build, train and deploy machine learning models. These capabilities are therefore essential for MLaaS-Platforms and define MLaaSP as platforms, that provide the ability to build, train and deploy machine learning models.

2.2.2 Applications of Machine Learning as a Service. Tang and Tay [55] applied deep learning algorithms within the area of wood identification to create an "Artificial Intelligence as a Service" system. The authors trained their model with macroscopic wood anatomy images and information about type and origin of the wood with the aim to fight illegal logging by harming the black market for illegally logged wood. Mariani et al. [27] present an architecture for a so-called Decision Support System, which aims to help doctors and other clinicians to determine the health risk of patients. The authors title this service Risk Predication as a Service. Furthermore, Mariani et al. [27] highlight the potential of MLaaS to reach higher effectiveness within model deployment while considering the risks regarding data privacy and data disclosure, which is especially relevant in the healthcare industry. Naous et al. [32] suggest applying MLaaS in their new defined paradigm Big Data as a Service. This paradigm covers infrastructure, data management resources, e.g., Hadoop, and advanced analytical capabilities, including ML algorithms. Other areas of application that we identified during our systematic literature research are urban modelling [30], computer chip development [58], personalized chatbot services [51], big data applications [38], Bug Prediction as a Service [53] and the testing of ML algorithms with MLaaS platforms as environment [37].

Nearly all of the applications that we identified during the literature review are constructed within the platform setting of MLaaS. A different approach is chosen by Pohl et al. [34]. The authors expand the existing service paradigm of providing only a platform or infrastructure to a Data Science as a Service (DSaaS) model. For the authors, data science comprises the provision, preparation, analysis, and visualization of data [34]. In conclusion, Pohl et al. argue that "providing a platform or a software that allows to conduct data science is not Data-Science-as-a-Service, however such an offer is not possible without. Data-Science-as-a-Service is a symbiosis of infrastructure, platform, software, and the processing of data science tasks" ([34] p. 437). We argue that this is equally true for ML: Providing a platform or a software that allows to conduct machine learning is not Machine Learning as a Service, however such an offer is not possible without. Machine Learning as a Service is a symbiosis of infrastructure, platform, software, modeling capabilities and its integration. Due to the fact, that this definition is not generally accepted within the scientific literature, we call this holistic view *Machine Learning Services* (MLS). To be independent from the platform setting, we define MLS as: "providing services along the complete project lifecycle from understanding the business needs to solving these via the application and deployment of machine learning models". The difference between MLS and MLaaSP can be visualized by the data science lifecycle from subsection 2.1.2.

The area in which the MLaaS platforms provide services in is indicated in the blue circle in figure 2. It is evident that the first step, Business Understanding, cannot be covered by the platforms as it comprises understanding and specifying the business problem as well as identifying relevant data sources. The second step, Data Acquisition & Understanding is only half covered by the MLaaS platforms. By this, the acquisition of data sources that are not yet connected to the platform environment is accounted for. The ability to build and train machine learning models is essential within the MlaaS platforms. Hence the Modeling step is completely covered in the visualization. By leveraging the provision of computing resources for the resource intense modeling, the platform providers utilize this step as their key service. Deployment is only half covered by the MLaaS platforms. While they allow for easy deployment within their own ecosystems, the model results still need to be made available to the consuming application. The last step, customer acceptance is of project-based nature and therefore not included in the MLaaS platform's offering.

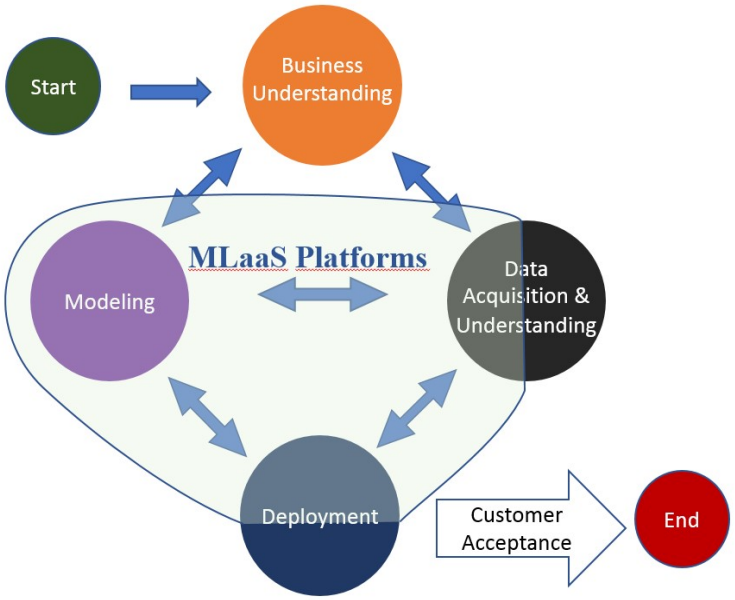

Figure 2 MLaaSP in the context of the Team Data Science Lifecycle (adapted from [29]).

2.2.3 Performance Enhancements for Machine Learning as a Service. Other than in [33] we refer briefly to the concept of performance enhancements in this contribution. Hazelwood et al. [14] mention substantial performance increases for certain use cases by training on GPUs rather than on CPUs. Training on GPUs offers shorter iteration times, and therefore enables the users to analyse either a higher number of ideas during a predefined period, or the same number of ideas in less time. Yao et al. (2017) concluded, that more user control may lead to higher performance in the hands of experienced users. Furthermore, the authors identified classifier choice as configure choice with the highest performance gains and that exploring the results of a short random subset of classifiers can lead to close-to-optimal outcomes. Liu et al. [26] come to a similar result. Comparing the MLaaS platforms of Microsoft Azure and Amazon Web Services (AWS) they conclude that the additional alternative models provided by Azure lead to a stronger performance within their data sets. A thoroughly performed selection of the presented models is essential as their results may vary considerably. The trade-off between accuracy and latency is analysed by Halpern et al. [13]. Different interests between the requirements of API consumers and the deployment of machine learning models in the cloud motivated the authors to introduce Tolerance Tiers. These tiers should enable consumers of such services to configure the API based on their operational requirements, e.g., cost, accuracy or responsiveness. Three different tiers are proposed, one to represent the base settings, one to minimize service response time, and one to minimize service costs. The authors show that their proposed Tolerance Tiers can optimize accuracy without decreasing responsiveness or increasing costs. Qin et al. [36] highlight existing ML systems reliance on manual parallelism configurations. The sequential manner, in which requests are usually executed can result in high latency, but common Service Level Objectives (SLO) demand between 500 and 800ms. To reach latencies this low, parallel computation is used. The authors propose a framework to schedule machine learning offerings and find the best configuration under changing workloads.

This subsection presented different ways to enhance the performance of both, machine learning models and MLaaS platforms. Technical solutions proposed are the use of GPU's, different API tolerance tiers, and a scheduling framework. Flexibility in alternating between different models and the application of ensemble models are advantages, that independent machine learning approaches have over MLaaS platforms.

2.2.4 Challenges within MLaaS and Machine Learning Services. In this subsection we recap the challenges within the area of MLaaS and MLS. Although the resources are solely based on our previously conducted literature review of MLaaS contributions, some of the challenges can be relevant for both. Table 1 shows the challenges of the different publications grouped into Lack of tools, Privacy issues, Prediction API threats, Heterogeneity of languages and environments, Lack of technical expertise and Ethical challenges. For the following section we focussed on those challenges and contributions that we classified as relevant for this study.

Contribution	Year	**Challenges**					
		Lack of tools	Privacy issues	Prediction API threats	Heterogeneity of languages	Technical expertise	Ethical Challenges
Ribeiro et al.	2016	X				X	
Zhang et al.	2016	X					
Hesamifard et al.	2018		X				
Hunt et al.	2018		X				
Kesarwani et al.	2018			X			
Masuda et al.	2018						
Agrawal et al.	2019				X		
Hesamifard et al.	2019		X				
Hitaj et al.	2019			X			
Hou et al.	2019			X			
Mariani et al.	2019				X		
Rao et al.	2019					X	
Reith et al.	2019			X			
Stoyanovich	2019						X

Table 2. Challenges in MLaaS and in MLS identified through a literature review [33].

The lack of technical expertise and the lack of tools are challenges that led to the emergence of MLaaS systems [42, 39]. Rao et al. [39] mention lack of expertise in implementing ML applications, while Zhang et al. [65] highlight the need for a tool that is capable of coordinating the complex interactions between code, data and other parameters. Model versioning is among these complex interactions and another challenge within MLaaS [27, 65]. Having high quality metadata and storing it separately is another issue [27]. Metadata is data, that provides information about other data with the purpose of defining the data [48]. It is crucial in ensuring the interpretability of both input and results.

Privacy is one of the biggest and most disputable challenge within MLaaS. Since machine learning algorithms need access to the raw, often privacy sensitive, data, it is challenging to ensure the security and privacy when training and deployment are done on cloud providers [16, 21]. Additional risks arise once the trained models are deployed on the cloud. Within MLaaS cloud-based platforms, the security and privacy of a trained model can be at risk to subversion attacks [23, 18, 20, 41]. There are three different types of subversion attacks. An evasion attack causes a model to misclassify by making changes to the input data [23]. The attacker wants to input data into a trained algorithm to produce

incorrect output [41]. A poisoning attack attempts to modify the training data in order to change model results after updates. Kesarwani et al. [23] define an extraction attack as the abuse of the query API of a model. This abuse enables the launch of smart queries in an attempt to steal the hosted model. Model confidentiality, privacy, and model revenue are at risk of those attacks. These attacks can further obtain access to model parameters that may include private information about the training data and the model. Certain sectors such as IT-security, banking or healthcare are especially at risk of these attacks due to their critical data [27].

Another big challenge in the field of MLaaS and MLS is the heterogeneity of languages [22]. To overcome this barrier, two standards have been developed. One is the Portable Format for Analytics (PFA) and the other one is the Predictive Model Markup Language (PMML). PFA "is a common language to help smooth the transition from development to production" [6]. Figure 3 illustrates how PFA can serve as an interface between more flexible languages such as Python and R, in which a lot of the ML models are written in, and other languages. The right side of the dashed line displays languages of developing tools that need to communicate with production environments, for example client-side web browsers. PFA abstracts the ML algorithms into PFA documents in JSON format, that serve as configuration files. This decoupling is essential because statistical models generally change more quickly than the data pipeline [6]. The PMML then standardizes analytical models further into an XML file and therefore enables sharing of trained models and their descriptions [27]. The challenge of heterogeneity of languages, data layouts, and formats gets more complicated, since data is often spread out into different silos within the businesses [2]. Data silos are collections of information within an organization that are inaccessible by other parts of the organization. This further affects securing data provenance, versioning, and compliance standards. A lack of communication within businesses can sometimes lead to multiple teams working on related ML problems [2].

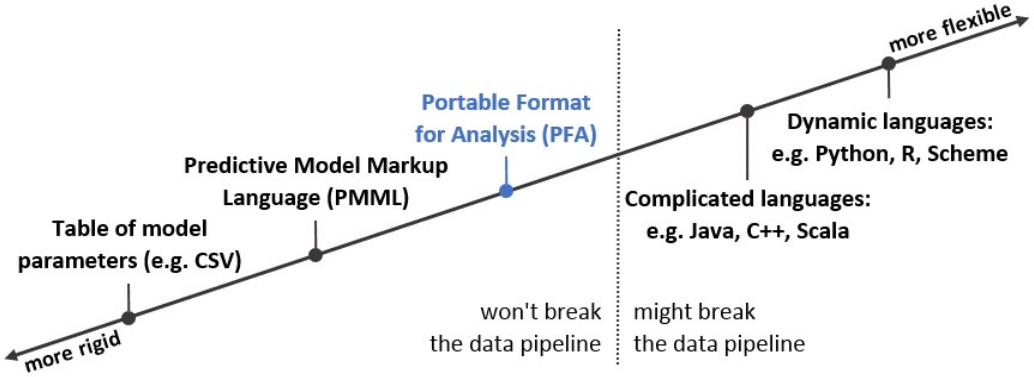

Figure 3 Language comparison ([27] p.302).

3. Methodological Approach and Delphi Study Design

After presenting the summarized results of our previously conducted literature review, this section first describes the employed method, sample, and procedure. Afterwards, we present the development of a Web-Application Tool for the Delphi Study as well as the criteria for the evaluation of the results.

3.1 Method, Sample and Procedure

In order to identify challenges and opportunities in the application of MLS, we conducted a Delphi study with a panel of machine learning experts. The objective of this method is that the panel of experts reaches a consensus among their formulated opinions regarding the challenges and opportunities. According to Rowe and Wright [44] the key features of the Delphi method are the anonymity of the participants, iteration for consensus generation and controlled feedback to improve and debias individual judgements. Additionally, an equal weighting of the experts allows for a statistical aggregation of group responses, so that quantitative analyses can be performed.

The Delphi method was chosen since MLaaS research is still novel, which our literature research showed. The exploratory nature of the classic Delphi method with open questions in the first round is well suited for this research. It might generate predictions that not just reflect the existing discussion in the scientific literature, but possibly reveal new aspects of the field. The method was applied to a similar research question of identifying risk factors within software projects by Schmidt et al. [47]. It was further applied to identify traits of top performing software developers by Wynekoop and Walz [62]. The different point of views of the different roles in machine learning projects should generate a diverse mix of responses. In this study, we used the multi-step framework presented by Skulmoski et al. [50], which is displayed in figure 4, exemplary for a three-round process.

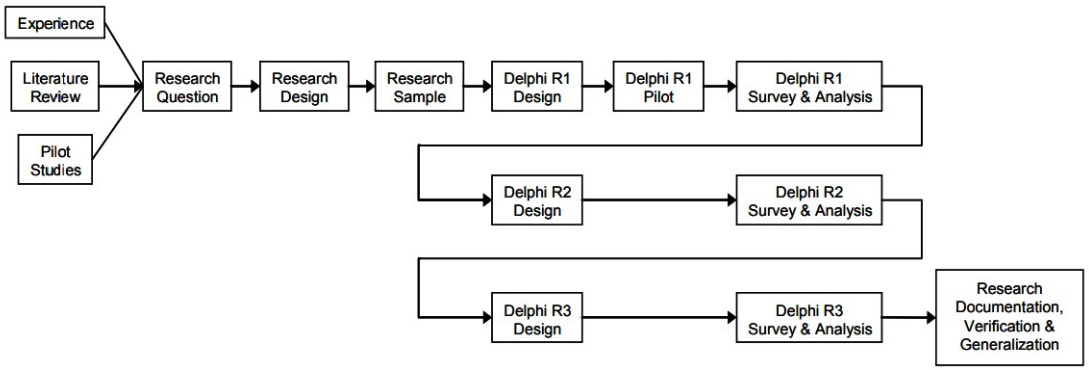

Figure 4 Three-round Delphi process ([50] p.3).

The research design of our study is a modification of Schmidt's method (cf. [46]), which uses nonparametric statistical techniques to analyse and report results in a ranking type Delphi survey. The amount of different challenges was initially expected to be much lower than the actual response of the experts. Due to the high number of named items, Schmidt's nonparametric ranking type analysis was not applicable. Instead, the respondents are asked in the second and third round to determine, how much they agree with the named challenges and opportunities on a five-point Likert-scale (1: Strongly Disagree to 5: Strongly Agree).

The selection of participants is crucial for the success of the method since the whole output is formed through the expert's opinion. To gather the respondents for our Delphi study we contacted various experts through the LinkedIn network and through emails to large actors in the industry. The participants consist of data scientists, solution architects and tech executives. Out of 30 contacted

experts, 14 respondents confirmed the study and 10 experts progressed until the end of round three. This approach is compliant with the literature, since the panel of experts is not selected to represent a general population, but rather to generate expert opinion [11; 4]. Our panel comprises experts from cloud computing platforms, the banking industry, IT consulting firms, Innovation consulting firms, AI startups and freelance data scientists.

Following Schmidt [46], we divided the survey process into three rounds of two different phases. Phase one covers the discovery of issues, where respondents should submit as many issues as possible, which is why we are asking open questions. For a comprehensive analysis, the questions were further divided into the different steps of the Data Science Lifecycle (cf. section 2.1.). Structuring the questionnaire into different steps allows for an easier mapping of similar answers onto the corresponding challenges. Furthermore, it helps the research panel structure their responses. The questionnaire of round one included a short introduction to the study, which comprised the definition for MLS as well as the figure of the Data Science Lifecycle. In the first round of the Delphi study, the experts named 162 challenges and 23 opportunities in total, which are coded onto 90 different challenges and 14 opportunities. To improve the reliability of the results, continuous verification and affirmation are crucial [1]. In phase two, the named factors from the first phase are randomly ordered and send out to each of the participants. The panellists need to reduce the list of factors to a more workable number of factors. This is done by presenting the participants the 104 named challenges and opportunities and asking how much they agree with each on a five-point Likert scale. An additional N/A option was given, which was to be selected if the participants had no opinion on or not enough knowledge about the item. We included this option due to some items being very specific for certain uses cases or mentioning specific technologies. According to [46], a second round of phase two needs to be conducted in case more than 20 items (in our case per step of Data Science Lifecycle) are remaining. Within our study, we excluded items with a mean rating lower than 3 or with more than 29% of N/A ratings. This was the case for 15 out of 104 named items, so that 89 items remained for the third round of our Delphi study. The specific deviation onto each step and the average rating in round two can be seen in table 2.

	Number of items in R2	Remaining items for R3	Average rating R2
Business Understanding	15	13	3,9
Data Understanding	15	14	3,7
Modeling	24	18	3,7
Deployment	20	17	3,6
Customer Acceptance	16	14	3,7
Opportunities	14	13	3,8
SUM	104	89	3,7

Table 2 Reduction of items for round three and average rating in round two.

The second round of phase 2 (or round three overall) contains information about the groups ranking of the different challenges and opportunities. The average ranking for each of the remaining 89 items is given to the participants within the online questionnaire. In addition, the respondent's own answer from round two is indicated for each item with the option to change his/her rating.

3.2 Developing a Web-Application as a Tool for the Delphi Study

For our study, we developed a Web Application and data pipeline to perform the Delphi study online. The motivation for such a resource-intense step was threefold: (1) complete control and freedom over the design of the survey, (2) ensuring data security for the entire study, and (3) signaling to the data science experts our competence, effort and know-how in the field. The online tool was realized utilizing the AWS Cloud, through a combination of personalized web interfaces communication via API endpoints with a backend consisting of our business logic and persistence layer. The persistence layer comprises the database to store the information. To make our data comparable, an extract process was developed, that transforms the data and loads it into Excel-Files. Figure 5 visualizes the architecture of the implemented solution.

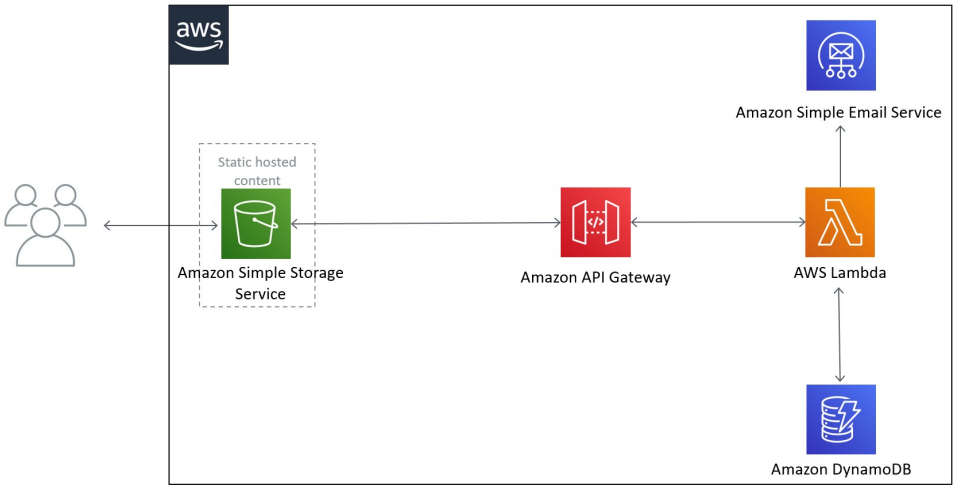

Figure 5 Architecture of the implemented tool for an online Delphi study.

3.3 Criteria for Evaluation

The evaluation of the results follows the procedure as suggested by Dajani et al. [7], who analysed stability and agreement criteria within Delphi studies and defined the following levels of agreement:

Consensus: Unanimity is achieved
Majority: More than 50 % of participants exhibit consistency
Bipolarity: Participants are equally divided; when Bipolarity occurs, the stability among the two bipolar groups should be assessed
Plurality: Less than 50 % exhibit consistency
Disagreement: Participants maintain independent views

For setting the necessary thresholds we followed von Briel [57] where consensus is assumed when more than 75% of respondents chose a rating of 4 or 5 on the Likert scale. Items with more than 50% but less than 75 % rating of either 4 or 5, are interpreted as a majority agreement. For the calculation of the percentages, the N/A option is interpreted as a missing value. The percentages are calculated for the number of respondents, who responded to an item with an actual rating. Those items that show

neither a consensus nor a majority agreement, were analysed further for bipolarity. Bipolarity means that respondents have two opposite positions on an issue [7]. Following [40] we interpret those items as possibly bipolar, where the interquartile range (IQR) of these items has a value greater than 1. The histogram analysis gives transparency of the extent of discrepancy amongst the experts' assessments, with two high peaks indicating bipolarity [59]. Items, that show Plurality or Disagreement by either an average rating lower than 3 or by more than 30% of N/A answers, were excluded after round two of the study. In round three, no items moved from stable to plurality or disagreement. Stability is assessed by the coefficient of variation (CV), which is the ratio of standard deviation and mean value [7]. The criteria for agreement that were used in this study are summarized in table 3.

Criteria	Measurement
Consensus	>75 % of the participants rank the item with 4 or 5
Majority Agreement	50 – 75 % of the participants rank the item with 4 or 5
Bipolarity	IQR > 1 and two-peak formation visible in histogram

Table 3 Criteria for the level of agreement.

4. Results

In this section, we present a detailed evaluation of results of the study, which are divided according to the steps of the data science lifecycle. In general, our study led to a higher consensus between the experts from round two to round three. In round three, every item showed an average CV below 0.5, which indicates a good degree of consensus and makes an additional round unnecessary [8]. To further analyse the different challenges over the whole data science lifecycle, we grouped them into seven areas. Table 4 lists the formed areas, ranked by number of mentions regardless of agreement level.

Area	Count
Communication	17
Environment	17
Approach	14
Model Training and Evaluation	9
Retraining, Testing, Monitoring and Updating	8
Data Identification/Acquisition	7
Data Quality	4

Table 4 Identified areas of challenges ranked by number of mentions.

4.1. Analysis of the Challenges in the individual steps in the TDSL

We structured the questionnaire according the different steps of the Team Data Science Lifecycle to support the mapping of similar answers onto the corresponding challenges. The following subsection elaborate on the results of each steps in detail.

4.1.1 Business Understanding. The participants of our study ranked 13 challenges in the Business understanding step of the data science lifecycle (Tab. 5). Two challenges are attributed towards the area of Approach, six towards Communication and four towards experience. A consensus was reached for the challenges *Clear problem definition*, *Unrealistic expectations from Business - Overexpectation in ML capabilities and underexpectation of actual implementation work*, *Companies treat ML Projects as IT Projects, Business expert's lack of time to invest additional resources into ML project, High amount of domain knowledge needed,* and *Bilateral knowledge transfer needed*, named in order of their mean rating.

Challenge	Sector	Mean Rating	Rating 4 or 5
Clear problem definition	Communication	4,7	100%
Unrealistic expectations from Business - Overexpectation in ML capabilities and underexpectation of actual implementation work	Communication	4,6	90%
Companies treat ML Projects as IT Projects	Approach	4,4	90%
Business expert's lack of time to invest additional resources into ML project	Approach	4,1	80%
Language gap between Business Experts and Data Scientists	Communication	4	70%
Technology First Approach - Evaluate if ML is needed to solve the problem	Communication	3,9	70%
High communication needed	Communication	3,8	60%
High amount of domain knowledge needed	Communication	3,8	80%
Bilateral knowledge transfer needed	Communication	3,75	75%
Selection of the metrics and performance criteria according to Business understanding	Approach	3,7	60%
Lack of experience with different ML approaches, bias towards approach with experience in	Approach	3,4	50%
Lack of talent at client side hinders the identification, specification, sourcing and implementation process	Approach	3,33	44%
Finding the relevant Business experts within an organization	Communication	3,2	40%

Table 5 Mean rating and consensus of the challenges within Business Understanding.

All three items without consensus or majority had an interquartile range higher than one. As their histograms did not show a two-peak formation, bipolarity cannot be concluded. *Clear problem definition* is the highest rated challenge over all parts of the lifecycle. It was further described by participants as: “Lack of clarity regarding the problem: the business cannot explain the problem, either because they don’t understand it very well or because they bring a solution they think should be implemented, instead of focusing on explaining the problem in detail” and “Understanding the business problem. For the data scientist, it is very challenging to understand the business problem well enough to adequately define the data science problem. This is not a problem for the work itself, which always is highly iterative and of an exploratory nature (at least in the beginning), but it is sometimes

hard for the business people to understand that mode of working." Participants further highlight communication issues between the business experts and the data scientists. These issues become visible within other named challenges as well. The *language gap between business experts and data scientists* and the *high amount of needed communication* to reach the *bilateral knowledge transfer* enhances the challenge of communication. *Unrealistic expectations from Business* is the second highest-rated challenge of all steps of the study. This is due to a lack of understanding of machine learning from the business side, and too high expectations toward the technology.

Overall, five of the challenges were grouped into the area approach. These primarily addresses problems with the conceptualization of ML project. The challenges can further be fine-structured into challenges on the business side and challenges on the machine learning expert's side. The experts highlight, that *Companies treat ML Projects as IT Projects*, which again is due to a poor understanding of machine learning on the business side. Another challenge, that can come with the poor understanding is that *Business expert's lack (of) time to invest additional resources into ML project*. Challenges for the ML experts are to *Evaluate if ML is needed to solve the problem,* and to have a *High amount of domain knowledge.* This knowledge is necessary to select the proper metrics and performance criteria. The challenge: *Lack of experience with different ML approaches, bias towards approach with experience in,* is further named in the third step modeling. The high amount of different complex techniques and methods makes it difficult to be experienced in all of them, and to select the best approach for each project. On average, the challenges within Business Understanding were rated the highest out of all lifecycle steps.

4.1.2 Data Understanding. In the data science lifecycle, the data understanding step comprises data source, pipeline, environment and wrangling, exploration and cleaning. Out of 14 challenges the panelists reached a consensus on 4 challenges, and a majority agreement on 7 (see table 6). The highest rated challenge is *Identifying the relevant existing and possible new data sources for the problem*. In round one already, this challenge was named by four experts. Second highest rated is the *Lack of or bad Metadata*. This challenge was further described by participants as: "data governance and management are usually not the top priority until a business need is identified. So, even when data was collected, descriptive information, referential integrity information, changes in systems (which often bring changes in data domains and break time series) and business shifts are not documented and a lot of the effort is spent trying to make sense of data points by just looking at numbers." The expert's opinions show that so far, metadata was not of high enough interest for businesses. This is again, due to a lack of understanding for the domain machine learning in general and the prerequisites for its application. The third highest rated challenge *Communicating the required data preparation workload understandably within proof-of-concept projects* is another communication issue, that is related to the lack of conceptual understanding on the business side. The challenge was further described as "The fact that data preparation is a lot of work is hard to communicate in exploratory proof-of-concept projects. Customers have a hard time understanding that to find out if a business problem can be solved with ML, the data pipeline needs to be built up adequately (even for a fail-fast approach). And they have a hard time to accept that this takes a long time (that 80% of time for data preparation and data understanding that everyone talks about, but still no one seems to accept as the basis for all ML/DS projects)." On fourth position is the challenge of *data being split across different silos*. This challenge could be attributed towards both areas, Data Identification & Acquisition, and Environment. Most companies do not have the infrastructure to enable efficient data identification and acquisition.

The lack of infrastructure is further mentioned in the challenge *Data Infrastructure not ready for ML-workflows*. The lack of data lakes and efficient data pipelines is mentioned by multiple experts. Data Quality is a generally accepted challenge. The experts named the lack of metadata, low data quality in general and Gaps in the existing data. Within the area Data Identification & Acquisition, the named challenges show, that a precise understanding of the data sources is crucial. Without this understanding, the data cannot be interpreted correctly.

Challenge	Sector	Mean Rating	Rating 4 or 5
Identifying the relevant existing and possible new data sources for the problem	Data Ident. & -Acqu.	4,3	90%
Lack of or bad Metadata	Data Quality	4,3	80%
Communicating the required data preparation workload understandably within proof-of-concept projects	Communication	4,11	78%
Data is split across different silos	Data Ident. & -Acqu.	4,1	80%
Comply with data protection obligations	Data Ident. & -Acqu.	3,9	70%
Low Data Quality	Data Quality	3,9	70%
Getting all required accesses	Data Ident. & -Acqu.	3,8	70%
Gain exact understanding about data Acquisition, to interpret missing values and outliers correctly	Data Ident. & -Acqu.	3,7	70%
Data Infrastructure not ready for ML-workflows	Environment	3,67	56%
Gaps in the existing data	Data Quality	3,6	60%
Making sure that models get continuously retrained on newly arriving data	Environment	3,56	56%
Inhomogenous or Legacy data formats	Data Quality	3,2	40%
Low amount of data	Data Ident. & -Acqu.	3,2	40%
Separate the required information from unnecessary inputs	Data Ident. & -Acqu.	3	30%

Table 6 Mean rating and consensus of the challenges within Data Understanding.

4.1.3 Modeling. The third step of the data science lifecycle covers feature engineering, model training and model evaluation. Table 7 shows the mean rating and the consensus and majority agreements. Four of the challenges had an interquartile range higher than one. The histograms showed no two-peak formation, bipolarity is therefore not present. A consensus was reached for five challenges. The first one is a data related challenge, namely The ML model results can only be as good as the data that the algorithm was fed. This challenge was further described as “If the data contains a lot of noise, the model quality won’t be good. Even worse, if the data contains bias, the model results will be biased too. And that bias is hard to define and hard to check for.” Second highest is another communication challenge, namely *Presenting the model results in a way that is understandable for the*

business experts and stakeholders. Another high rated communication issue shows, how important it is to enable the business experts and stakeholders to understand the approach of the model and its results.

Challenge	Sector	Mean rating	Rating 5 or 4
The ML model results can only be as good as the data that the algorithm was fed	Model Training and Evaluation	4,4	90%
Presenting the model results in a way that is understandable for the business experts and stakeholders	Communication	4,3	90%
Enough business understanding to guide feature engineering and model development	Approach	4,2	90%
Define appropriate target/success metric	Model Training and Evaluation	4,2	90%
Carefully select validation and test set, satisfactory criteria and optimization metric	Model Training and Evaluation	4,1	70%
Finding a good balance between model interpretability and model performance	Model Training and Evaluation	3,9	70%
Time and budget limitations	Approach	3,89	78%
Having enough data points to build robust models	Model Training and Evaluation	3,8	70%
Imbalanced representation of classes	Model Training and Evaluation	3,7	70%
Structured approach needed due to too many options for ML-methods	Approach	3,67	44%
Achieving high generalisation capabilities	Model Training and Evaluation	3,56	44%
Finding a good balance between a quick-and-dirty model to check if a use case is feasibly to be solved with machine learning and getting the best possible model	Approach	3,5	60%
Decreased productivity due to significant additional effort required for dataset/model parameter versioning to ensure reproducibility	Approach	3,5	40%
Lack of experience with different algorithms, bias towards understood algorithms	Approach	3,4	40%
Performant, easy to use modeling environment	Environment	3,33	44%
Meaningful Dimensionality reduction	Model Training and Evaluation	3,3	30%
Lack of capacity and compute resources at customer side push most of the workload outside the company, which leads to a lack of inhouse know-how	Environment	3,2	40%
Nobody noticing poor quality because nobody understands what the data scientist does	Model Training and Evaluation	3,1	50%

Table 7 Mean rating and consensus of the challenges within Modeling.

Most of the challenges in this step are related to Model Training and Evaluation. The two highest rated related challenges deal with the clear definition of the target-/success metric and validation- and test set. Clear definitions are crucial for a data scientists work, especially for ML projects. These challenges extend the highest rated challenge from the business understanding sector, *Clear problem definition*, into the modeling phase where the clear problem definition needs to be translated into clear definition of the key metrics. Two challenges with a majority agreement deal *with finding a good balance between (1) model interpretability and model performance* and *(2) a quick-and-dirty model to check if a use case is feasibly to be solved with machine learning and getting the best possible model.* The first challenge is further described as "For some purposes it is more important to understand why a model predicts what it does, and for other the prediction quality matters more, and it is okay for the model to be a black box. Talking to the businesspeople is important to find out what the model is needed for. It is challenging to find a common language for taking that decision." The second one highlights the exploratory nature of the approach. Only by analyzing the given data, the data scientist can determine if the problem should be solved with machine learning.

4.1.4 Deployment. The deployment step connects the model with the environment, in which the model's predictions or results are applied to. Hence a high number of challenges deal with the sector environment. All of the challenges are displayed in table 8. Highest rated is the *Lack of experience and processes for productive model management*. The panelists further highlighted the lack of "awareness of the need for, and ability to implement processes for testing, versioning, retraining, change management when data sources change, monitoring etc." within that challenge. This lack can be partially accredited to the *Gap between modeling and deployment*, the second highest rated challenge. "Having to reimplement the model and/or data pipeline in another language for deployment. It is a big challenge when models have to be reimplemented in a different language or environment for being productionalized" highlights the difference between the modeling and deployment environment as a big challenge. This difference and the lack of experience with deployment environment contribute to the fact that *Robustness, monitoring, speed and maintainability concerns are only taken into account during deployment*, the third highest rated challenge. *Organization is not ready for deploying* is the last challenge, that the panelists found a consensus on. This challenge is present in other steps as well, e.g., Data Infrastructure not ready for ML-workflows. The experts state that in general, the business's infrastructure is often not ready for the ML projects. This lack of infrastructure affects multiple steps of the project's lifecycle.

One panelist stated that a "lot of organizations try some kind of ML project, and after a successful proof of concept have a hard time to continue to move forward with their ML endeavors. Often, the organization is not quite ready: The data is too decentralized, there is no structure and / or infrastructure to form a data science team, or no one wants to make that decision to carry on." Five of the challenges were related to Retraining, Testing, Monitoring and Updating. A consensus was reached on Automation of quality assurance of input and output data. A panelist further stated that a "proper deployment requires a clear definition of the business processes around the actual ML-Service. These services must cover the quality assurance of both input and output data. Any recurrent scoring ML-process requires automation of quality assurance of input and output data. It also requires recurrent model evaluation and ML-outcomes (ML-scores, ML-generated business insights, etc.). The application data model needs to be extended with historization of model scores and historized data around business outcomes generated by using ML-scores." The description highlights the need for

clear definitions of the business processes and the input- and output-data to ensure their quality and therefore the quality of the ML model's output. Other challenges with a majority agreement attributed to Retraining, Testing, Monitoring and Updating are integrating performance measurement into deployment, Securing interpretability to be compliant with data protection obligations and Model update requires large efforts.

Challenge	Sector	Mean rating	Rating 5 or 4
Lack of experience and processes for productive model management	Environment	4,33	89%
Gap between modelling and deployment environment requires redevelopment	Environment	4,11	89%
Robustness, monitoring, speed and maintainability concerns are only taken into account during deployment	Approach	4,11	78%
Automation of quality assurance of input and output data	Retraining, Testing, Monitoring and Updating	4,1	80%
Organization is not ready for deploying	Environment	4	89%
Proper deployment requires a clear definition of the business processes around the actual ML-Service	Approach	3,9	70%
Underestimated costs for running live inference for a long period of time in production and costs for collecting and managing the data required to retrain the model and keep it up-to-date are	Environment	3,78	67%
Integrating performance measurement into deployment	Retraining, Testing, Monitoring and Updating	3,67	56%
Having the right architecture	Environment	3,6	60%
Securing interpretability to be compliant with data protection obligations	Retraining, Testing, Monitoring and Updating	3,56	56%
Model update requires large efforts	Retraining, Testing, Monitoring and Updating	3,44	56%
Costly infrastructure required to run, monitor and scale the applications automatically	Environment	3,4	40%
Defining responsibilities, timeframe, manual/automatic for retraining	Retraining, Testing, Monitoring and Updating	3,4	30%
Scale of deployment from models, trained on thousands of GPUs/TPUs, creates new challenges	Environment	3,33	22%
APIs and data formats need to be well defined	Environment	3,33	44%
Low degree of knowledge regarding cloud provider security	Environment	3,3	50%
Complicated setup of deployment solutions in the cloud and on-premise	Environment	3,3	40%

Table 8 Mean rating and consensus of the challenges within Deployment.

4.1.5 Customer Acceptance. The final step of the lifecycle, customer acceptance, covers finalizing the project deliverables. Amongst other objectives it includes confirming that data pipeline, model and the deployment into the production environment are according to the customers satisfaction. Sector, mean rating, consensus and majority agreement for all the challenges are displayed in table 9. The three consensus challenges all belong to the communication sector. All panelists agreed on the *Difficulty to explain many models (Black-box nature)*. This relates to the issue of explainability, named in step 1, especially when more complex algorithms like neural networks are used. Second highest

rated is *Expectation Management*. Here the panel highlighted how unrealistic expectations often lead to disappointment with the results. The business expectations should be set right at the beginning before the project start. 80 % of participants agreed on *Explaining the always existing certain amount of controllable but not avoidable error*. A panelist highlighted that "any ML-result is nothing else than statistics and that any statistical result includes" the certain amount of error. This is again, due to low understanding of the statistical methods on the business side and contributes to having too high expectations. Third highest rated was *ML applications that are not built with business KPIs attached and that are not tested and validated by the business will struggle to be adopted*. This challenge goes hand in hand with *clear problem definition* and *clear definitions for the success metrics*. Without continuous precision until the customer acceptance step, which includes testing and validation, the results of ML projects will be subpar.

Challenge	Sector	Mean rating	Rating 5 or 4
Difficulty to explain many models (Black-box nature)	Communication	4,4	100%
Expectation Management	Communication	4,22	78%
ML applications that are not built with business KPIs attached and that are not tested and validated by the business will struggle to be adopted.	Approach	4,11	67%
Explaining the always existing certain amount of controllable but not avoidable error	Communication	4,1	80%
Testing tends to be a late concern	Retraining, Testing, Monitoring and Updating	3,89	67%
Live performance monitoring is not well implemented	Retraining, Testing, Monitoring and Updating	3,8	70%
Proper Documentation	Communication	3,7	60%
Lack of understanding of testing methods leads to mistakes in experiment design	Retraining, Testing, Monitoring and Updating	3,67	67%
Ensuring knowledge transfer	Communication	3,67	56%
Changing business environment	Environment	3,67	44%
Unfounded data privacy concerns	Environment	3,67	67%
Black-box nature can lead to regulatory concerns	Environment	3,6	70%
Difficulty to convince end users of the ML solution's benefit	Communication	3,44	56%
Lack of understanding regarding the intended effects make it difficult to calculate businesscase and amortization for client	Communication	3,33	56%

Table 9 Mean rating and consensus of the challenges within Customer Acceptance.

4.2. Analysis of the Challenges over all steps in the TDSL

To summarize our results, we provide an overview where the identified challenges are associated with the different life cycle steps, and their deviation onto the different sectors is indicated by color (Fig. 6). Only challenges with either a consensus or a majority agreement are included and within the colored clouds, a higher position means a higher average rating by the panel. The challenges deviate onto the sectors as shown in table 10. The second column displays the number of all challenges that the experts could rank in round 3, while the third column indicates on how many of these challenges a consensus or majority agreement was reached upon. For the following analysis the areas Data Identification & Acquisition and Data Quality were merged into the area Data.

Area	All R3	Cons. Or Maj. Agr.
Communication	17	16
Environment	17	9
Approach	14	10
Data	11	8
Retraining, Testing, Monitoring and Updating	8	7
Model Training and Evaluation	9	6

Table 10 Areas of challenges with consensus and majority agreement.

The highest number of challenges are identified within the area of communication. Clear definitions, expectation management and making approach and results understandable to the business experts are among the most important challenges within MLS. The communication challenges are crucial due to (1) an often-existing language gap between the business- and machine-learning experts and (2) general lack of understanding for ML methods and statistical methods in general. This lack of understanding is also visible in the approach section. Companies treating ML projects as IT projects and not allocating enough resources to these projects are among the highest rated challenges within the

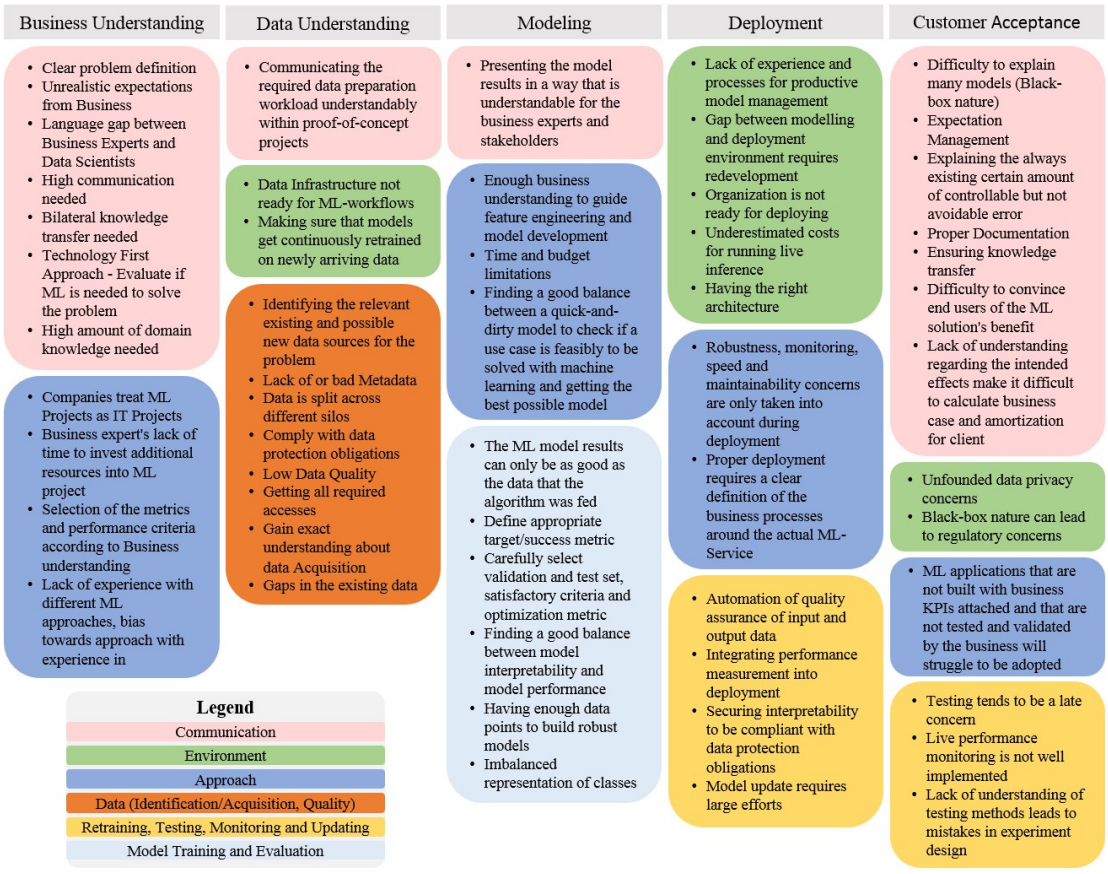

Figure 6 Overview of all challenges with a consensus or majority agreement by area.

Approach area. Other challenges in this area concern that robustness, monitoring, speed and maintainability are only too late considered and further the proper selection of metrics and performance criteria. These are of more interest for the machine-learning expert than for the businesses. Within the area Environment the gap between the modeling and deployment environment and the businesses lack of deployment ready infrastructure are among the highest rated challenges. The experts further agree on the challenge that within a lot of projects the businesses data infrastructure is not ready for ML workflows. These challenges show that the business's infrastructure is in general often not ready for ML projects. A lack of experience and well-defined processes for productive model management is further highlighted by the panel. An infrastructure assessment prior to the project start can identify possible infrastructure areas that demand improvement. General competence areas to be assessed are among others reporting, business intelligence and predictive analytics [35]. Lorentz [24] specified these infrastructure requirements: flexible and agile, efficient data management; can accommodate large and different volumes and data types; processing power and time; can receive data from different sources; manages the full data lifecycle and scales non-disruptively. The data area comprises both, Data Quality- and Data Identification- & Acquisition-challenges. Within Data Quality, a lack of or bad metadata is ranked highest, before low data quality in general and gaps in the existing data. These high ratings show the importance of proper metadata and data quality for machine-learning, Metadata is especially important to ensure interpretability and compliance with data protection regulations. Within Data Identification and Acquisition the panelists highlight the challenge of identifying the relevant data sources, which are often split across different data silos. This split is mostly due to different technologies being used within a business. They can further lead to inconsistency and a waste of resources. The different silos also lead to multiple different accesses that are needed. Getting these is another challenge the majority of panelists agreed upon. Within the Model Training and Evaluation area, highest rated is the challenge that the model results can only be as good as the input data and the definitions of appropriate target/success metric. Again, clear definitions and data quality are mentioned as big challenges. The last area is Retraining, Testing, Monitoring and Updating. The automation of quality assurance of input and output data is highest rated in this group.

4.3. Opportunities within the area of MLS

The last section of the Delphi study asked for opportunities within the area of MLS. Table 11 shows sector, mean rating, consensus and majority agreement for all opportunities that the panelists ranked in round three. Out of 13 opportunities, a consensus was reached for four and a majority agreement for six. All participants agreed on the opportunity to *Generate new business insights through data-driven approach.* The mentioned data-driven approach is seen as a chance to first get better insights by analyzing high quality data and then to translate these insights into data-driven decisions, which will lead to *Better business results by making important decisions based on vast amounts of empirical data instead of hierarchy.* By basing these decisions on data, and not on hierarchy or intuition, the decisions can then become *reproducible, understandable* and *comparable.* The panelists reached a consensus on the opportunity to *Apply ML to more projects in the realm of traditional industries.* Several different applications were shown in subsection 2.2.2 from wood identification, to risk prediction in healthcare as well as bug prediction. Further opportunities within the area Efficiency were named. These cover *cost savings*, *leveraging expertise*, *reducing costs of technology* and *levels of required skill*. The lowest rated opportunity mentions AutoML systems. AutoML stands for automated machine learning. These

systems aim at automating the whole ML workflow "including pre-processing, machine learning algorithm selection and hyperparameter optimization" [10]. Although our panel did not reach an agreement on this opportunity it is to note that AutoML systems are a promising area of research with multiple contributions and scientific challenges [12].

Opportunity	Sector	Mean rating	Rating 5 or 4
Generate new business insights through data-driven approach	Approach	4,4	100%
Better business results by making important decisions based on vast amounts of empirical data instead of hierarchy	Decision Making	4,22	89%
Apply ML to more projects in the realm of traditional industries	Applications	4,2	90%
Making decisions reproducible and understandable	Decision Making	4,1	90%
Cost savings and efficiency. Do more with less/same	Efficiency	3,9	70%
Leveraging expertise	Efficiency	3,9	60%
Surpass previous limits of automation with ML services	Applications	3,89	67%
Transfer Learning	Approach	3,7	60%
Reducing the barriers of adoption by constant lowering of cost of technology and level of skills required	Efficiency	3,6	60%
Make decisions comparable among each other by using the same model	Decision Making	3,44	44%
Producing even higher effects of scale	Efficiency	3,44	67%
Reinforcement Learning opens up ways to explore areas where data collection was impossible or limited to a small number of companies with large customer bases	Approach	3,2	30%
AutoML systems that are able to automatically analyze a dataset, implement automatic imputation, cleansing and engineering, pick adequate modeling approaches and fine tune algorithms	Efficiency	3	20%

Table 11 Mean rating and consensus of the opportunities within Machine Learning Services.

5. Discussion

The results of the Delphi study have revealed a variety of different challenges as well as opportunities. The highest number of challenges that the panelists found a consensus or majority agreement on belongs to the area of Communication. Furthermore, a general consensus exists about the importance of clear definitions and the existing language gap between business experts and data scientists. While these challenges are not new, the amount and the rating show its importance. Bridging the technical expertise of the ML specialists with the operational business expertise of the project's sector can be done by integrating a new role into the process, an analytics translator. The analytics translator can help in ensuring process success by making sure that the analytical models translate into operational impact [15]. Additionally, they allow the ML experts to focus on their key capabilities. Our results

show the need for either, the integration of this new role into the processes, or a greater focus by highlighting communication as the big challenge within ML projects. A general understanding of ML and the associated challenges can be achieved by incorporating knowledge transfers inside businesses and facilitating employee growth [5]. Brown et al. [5] propose an in-house analytics academy for large companies to further enable transformation towards data-driven business. Several environment challenges were named by the experts. In general, businesses are often not ready for ML projects and overestimate their readiness as well as the quality of their available infrastructure. This affects multiple steps of the lifecycle, as our results have shown. The existence of data silos within a business further obstructs ML projects and hinders them from unleashing their potential. Prior to the ML project, a comprehensive infrastructure assessment of the businesses technology systems and IT environment should be performed. Such assessment shall cover all systems that are affecting the project. Focus needs to be on the data infrastructure, to ensure that the models can get continuous quality data input. Another focus should be laid on retraining and monitoring to enable constant improvements of the model.

The results of the empirical study can partially be applied to MLaaS platforms. As shown in sub-section 2.2.1, MLaaS platform offerings comprise the lifecycle steps Data Understanding and Deployment partially, and Modeling completely. The challenges of the other two steps Business Understanding and Customer Acceptance are therefore equally important for the application of MLaaS platforms and MLS. The need for communication, clear definitions and sufficient time investment are exemplary challenges that are equally important for a company whether they are utilizing a MLaaS platform or utilizing internal or external ML experts. Additionally, the challenges Expectation Management and Explainability, named in the Customer Acceptance step, are equally important to both ML application models. While these challenges are of interest for users of MLaaS platforms, the challenges of the lifecycle steps Data Understanding, Modeling and Deployment are of interest for the platform provider. For example, an understandable presentation of the results, which was the second highest rated challenge in the Modeling step, should be provided through the MLaaS platform.

When comparing the challenges identified by the performed Delphi Study with the literature found in our previously conducted literature review, only some similarities could be found. First to note, the literature review was conducted aiming at the paradigm MLaaS while the Delphi study focused on our broader definition of Machine Leaning Services. The challenges within MLaaS platforms can only partially be compared with the challenges for ML services. Since the challenges, identified in the Delphi study come from multiple different areas, it would exceed the scope of this study to cover each area in a literature review. The identified communication challenges might be subject to scientific research, but not in context with MLaaS in its current definition. Since most of the communication challenges occur within Business Understanding and Customer Acceptance, these challenges are not covered by most of the MLaaS literature, as the two steps are not covered by MLaaS platforms. An often-addressed challenge in scientific literature are subversion attacks [23, 18, 20, 41]. This type of challenge was possibly not named by the ML experts since the definition of MLS was used in our study. MLS do not implicate sharing of the generated model. Hence the risk of subversion attacks is not present by default within this concept. Data being split across different silos is mentioned by [2] regarding a lack of integrated data platform solutions for ML that allow for the collaboration between different teams such as legal, compliance, quality control, ML engineering, etc. The lack of collaboration leads to a loss of consistency across different data silos, which may become the

bottleneck of ML projects. They leave the ML workflow exposed to physical data dependence on different teams and different data sources. The existence of silos makes "data sharing and discovery, data lineage and provenance tracking, version control, and access control difficult, if not impossible" [2]. The experts reached a consensus on this challenge. Another consensus was reached on the challenge of Lack of or bad Metadata. Zhang et al. [65] highlighted the need for separating data from their metadata, which leads to higher efficiency in ML workflows. Mariani et al. [27] show the importance of metadata by including metadata storage into their proposed decision support system. This allows the system to absorb already trained models and keep the interpretability. The authors further point out the challenge of heterogeneity of languages. This heterogeneity is included in the gap between modeling and deployment environment. The experts found a consensus on that challenge. Subsection 2.2.4 presented two standards, PFA and PMML as solutions for that issue. Ribeiro et al. [42] and Rao et al. [39] mention the lack of expertise in implementing ML applications as a contributing factor to the development of MLaaS platforms. While MLaaS platforms lower the need for expertise in the key areas Data Understanding, Modeling and Deployment, the lack of experience is still present and affecting the other two lifecycle steps. The lack of experience further contributes towards the businesses not being ready for deployment, since the prerequisites for ML projects are unclear to them. While the experts did not find a consensus or majority agreement on the issue of versioning, it is present in the scientific literature, with contributions made by Mariani et al. [27] and Zhang et al. [65]. Within the opportunities section of the study, the panellists found a consensus on the topic of applying ML to more traditional industries. Tang and Tay [55] have applied ML to wood identification by utilizing image recognition deep learning models, showing that research is being conducted into this area.

6. Conclusion and Future Work

The goal of our study was to identify challenges within the applications of Machine Learning Services (MLS). To reach this goal, we conducted a Delphi study. The empirical study identified 55 different challenges from six different areas. This high number of problems identified is a strong indicator for the complexity of ML projects. Over the entirety of the project lifecycle, challenges from multiple areas need to be overcome to successfully complete a ML project. The challenges we identified in our study belong to a wide variety of different sectors, namely Communication, Environment, Approach, Data (Identification, Acquisition, Quality), Retraining, Testing, Monitoring and Updating, and Model Training and Evaluation. The highest number of challenges where the panelists found a consensus, or majority agreement on, belongs to the area of Communication. The overall highest mean ratings belong to the importance of clear problem definitions, unrealistic expectations and the difficulty to explain models, all belonging to the area of Communication. Although these challenges are not new, the high ratings as well as the amount of named challenges, highlight the importance of maintaining effective communication in MLS. Furthermore, while the mean ratings were slightly lower for the Environment area, the high number of named challenges show the importance of this area as well. In general, most of these challenges originate from a lack of readiness or/and understanding of the technology. This often leads businesses to overestimate their existing knowledge and the quality of their available infrastructure. For this reason, the true potential of ML can often not be realized. Therefore, in our discussion, we explained how the assessment of the business infrastructure can help to increase the readiness and minimize these challenges.

For future work, it would be interesting to compare the challenges in MLS identified through our Delphi study with recent scientific literature available and identify possible conformities, nonconformities, and gaps. Because the literature review we had performed prior to our Delphi study had another purpose we were only able to compare challenges that were named within contributions regarding MLaaS. We leave further analysis of the challenges specific to ML as future work. Thus, we propose to either perform an explorative literature review, or to perform specific reviews for each of the six identified challenge areas. Overall, the arising complexity of ML projects over the complete lifecycle has not been researched on a broad basis. While our study identified many challenges within MLS, it would be another task for future work to distinguish those specific challenges that are unique to ML projects from those which generally appear in IT projects. In a next step, clear distinctions or interdependencies from one another and their impact may be investigated.

After all, the Delphi technique is subject to restrictions due to its qualitative and explorative character. Prior research regarding MLaaS first focused on the platform architectures and developed into different applications and challenges. Multiple studies analyzed different types of subversion attacks and appropriate countermeasures. More recent contributions focused on different applications of MLaaS platforms to enable new service models – another promising area for future research. There are many fields to which MLaaS platforms and ML can be applied to, which was further identified as an opportunity within the empirical study. The increasing complexity leads to the emergence of new roles within projects and businesses. An example being the business translator, which has been the focus of very little research thus far and could be the subject matter for later research. Additionally, we identified six further areas for research through the conducted Delphi study. The identified individual challenges can serve as starting points for future analysis. Additional contributions could add more application-oriented approaches that analyze multiple ML projects (e.g., in the form of case studies or best practice) and evaluate the impact of the different challenge areas on the projects' results.

Acknowledgements

The authors wish to express their gratitude to the panellists of the Delphi study. Research like this would not be possible without industry leading experts' contribution. We would further like to thank Simon Philipp for his guidance with the development of our Delphi Study Web Application.

References

1. Adler, M. and Ziglio, E., 1996. Gazing into the oracle: The Delphi method and its application to social policy and public health. Jessica Kingsley Publishers.
2. *Agrawal, P., Arya, R., Bindal, A., Bhatia, S., Gagneja, A., Godlewski, J., Low, Y., Muss, T., Paliwal, M.M., Raman, S., Shah, V., Shen, B., Sugden, L., Zhao, K. and Wu, M.C., 2019. Data Platform for Machine Learning. In Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19). ACM, New York, pp. 1803-1816.
3. Arute, F., Arya, K., Babbush, R., Bacon, D., Bardin, J.C., Barends, R., […] and Martinis, J.M., 2019. Quantum supremacy using a programmable superconducting processor. Nature 574, 7779, pp. 505-510.
4. Bourque, L. and Fielder, E. P., 2003. How to conduct self-administered and mail surveys. Vol. 3. Sage Publications.
5. Brown, S., Gandhi, D., Herring, L., and Puri, A., 2019. The analytics academy: Bridging the gap between human and artificial intelligence. The McKinsey Quarterly, September 2019, pp. 1-9.

6. Data Mining Group: Portable Format for Analytics (PFA) Retrieved from http://dmg.org/pfa/docs/motivation/, accessed January 10th 2021.
7. Dajani, J. S., Sincoff, M. Z. and Talley, W. K., 1979. Stability and agreement criteria for the termination of Delphi studies. Technological forecasting and social change, 13(1), pp. 83-90.
8. English, J. M. and Kernan, G. L., 1976. The prediction of air travel and aircraft technology to the year 2000 using the Delphi method. Transportation research, 10(1), pp. 1-8.
9. Farely, B. and Clark, W., 1954. Simulation of self-organizing systems by digital computer. Transactions of the IRE Professional Group in Information Theory, 4(4), pp. 76-84.
10. Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., and Hutter, F., 2018. Practical automated machine learning for the automl challenge 2018. In International Workshop on Automatic Machine Learning at ICML, pp. 1189-1232.
11. Fink, A. and Kosecoff, J., 1985. How to Conduct Surveys: A Step by Step Guide.
12. Guyon, I., Bennett, K., Cawley, G., Escalante, H. J., Escalera, S., Ho, T. K., ... and Viegas, E., 2015. Design of the 2015 chalearn automl challenge. In 2015 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1-8.
13. *Halpern, M., Boroujerdian, B., Mummert, T., Duesterwald, E. and Reddi V.J., 2019. One Size Does Not Fit All: Quantifying and Exposing the Accuracy-Latency Trade-Off in Machine Learning Cloud Service APIs via Tolerance Tiers. In 2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS). IEEE, pp. 34-47.
14. *Hazelwood, K., Bird, S., Brooks, D., Chintala, S., Diril, U., Dzhulgakov, D., […] and Wang, X., 2018. Applied machine learning at facebook: A datacenter infrastructure perspective. In Proceedings of the 24th IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, pp. 620-629.
15. Henke, N., Levine, J. and McInerney, P., 2018. You don't have to be a data scientist to fill this must-have analytics role. Harvard Business Review, February 5.
16. *Hesamifard, E., Takabi, H., Ghasemi, M. and Wright R.N., 2018. Privacy-preserving Machine Learning as a Service. Proceedings on Privacy Enhancing Technologies 2018(3), pp. 123-142.
17. *Hesamifard, E., Takabi, H., and Ghasemi, M., 2019. Deep Neural Networks Classification over Encrypted Data. In Proceedings of the 9th AMC Conference on Data and Application Security and Privacy (CODASPY'19). ACM, pp. 97-108.
18. *Hitaj, D., Hitaj, B. and Mancini, L.V., 2019. Evasion attacks against watermarking techniques found in mlaas systems. In Proceedings of the 6th International Conference on Software Defined Systems (SDS). IEEE, pp. 55-63.
19 Hoffman, D.L. and Novak, T., 2015. Emergent Experience and the Connected Consumer in the Smart Home Assemblage and the Internet of Things. SSRN Electronical Journal, pp. 1-151.
20. *Hou, J., Qian, J., Wang, Y., Li, X.-Y., Du, H. and Chen, L., 2019. ML defense: Against prediction API threats in cloud-based machine learning service. In Proceedings of the International Symposium on Quality of Service (IWQoS'19). IEEE/ACM, pp. 1-10.
21. *Hunt, T., Song, C., Shokri, R., Shmatikov, V. and Witchel, E., 2018. Chiron: Privacy-preserving Machine Learning as a Service.
22. Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349 (6245), pp. 255-260.
23 *Kesarwani, M., Mukhoty, B., Arya, V. and Mehta, S., 2018. Model Extraction Warning in MLaaS Paradigm. In Proceedings of the 34th Annual Computer Security Applications Conference (ACSAC'18). ACM, pp. 371-380.
24. Lorentz, C., 2018. Is Your IT Infrastructure Ready for Machine Learning and Artificial Intelligence. Retrieved from https://blog.netapp.com/is-your-it-infrastructure-ready-for-machine-learning-and-artificialintelligence/, accessed January 10th 2021.

25. Linnainmaa, S., 1976. Taylor expansion of the accumulated rounding error. BIT Numerical Mathematics, 16, pp. 146-160.
26 *Liu, Y., Zhang, H., Zeng, L., Wu, W. and Zhang, C., 2018. MLbench: Benchmarking machine learning services against human experts. Proceedings of the VLDB Endowment, 11(10), pp. 1220-1232.
27. *Mariani, S., Zambonelli, F., Tenyi, A., Cano, I. and Roca, J., 2019. Risk Prediction as a Service: A DSS Architecture Promoting Interoperability and Collaboration. In 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, pp. 300-305.
28. *Masuda, S., Ono, K., Yasue, T. and Hosokawa, N., 2018. A survey of software quality for machine learning applications. In Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM). IEEE, pp. 279-284.
29. Microsoft. 2020. The Team Data Science lifecycle. (Janurary 2020) Retrieved January 10, 2021 from https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle
30. *Milton, R. and Roumpani, F., 2019. Accelerating Urban Modelling Algorithms with Artificial Intelligence. In 2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). INSTICC, pp. 105-116.
31. Mladenow, A., Kryvinska, N. and Strauss, C., 2012. Towards cloud-centric service environments. Journal of Service Science Research, 4(2), pp. 213-234.
32. *Naous, D., Schwarz, J.S. and Legner, C., 2017. Analytics as a Service: Cloud Computing and the Trans-formation of Business Analytics Business Models and Ecosystems. In Proceedings of the 25th European Conference on Information Systems (ECIS). AIS, pp. 487-501.
33. Philipp, R., Mladenow, A., Strauss, C., and Voelz, A., 2020. Machine Learning as a Service – Challenges in Research and Applications. In Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services (iiWAS20). ACM, pp. 396 – 406.
34. *Pohl, M., Bosse, S. and Turowski, K., 2018. A Data-Science-as-a-Service Model. In Proceedings of the 8th International Conference on Cloud Computing and Services Science (CLOSER'18). SCITEPRESS, pp. 432-439.
35. Prout, A., 2019. Best practices for preparing your infrastructure for machine learning and AI. Retrieved from https://www.dig-in.com/opinion/best-practices-for-preparing-your-infrastructure-formachine-learning-and-ai, accessed January 10th 2021.
36. *Qin, H., Zawad, S., Zhou, Y., Yang, L., Zhao, D. and Yang, F., 2019. Swift machine learning model serving scheduling: a region based reinforcement learning approach. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'19). ACM, pp. 1-23.
37. *Rajagopal, S., Hareesha, K.S. and Kundapur, P.P., 2020. Performance analysis of binary and multiclass models using azure machine learning. International Journal of Electrical & Computer Engineering, 10(1), pp. 978-986.
38. *Ramachandran, M., 2019. SOSE4BD: Service-Oriented Software Engineering Framework for Big Data Applications. In Proceedings of the 4th International Conference on Internet of Things, Big Data and Security (IoTBS'19). SCITEPRESS, pp. 248-254
39. *Rao, S.S., Pradyumna, S., Kalambur, S. and Sitaram, D., 2019. Bodhisattva - Rapid Deployment of AI on Containers. In Proceedings of the 7th International Conference on Cloud Computing in Emerging Markets (CCEM). IEEE, pp. 100-104.
40. Rayens, M. K. and Hahn, E. J., 2000. Building consensus using the policy Delphi method. Policy, politics, & nursing practice, 1(4), pp. 308-315.

41. *Reith, R.N., Schneider, T. and Tkachenko, O., 2019. Efficiently Stealing your Machine Learning Models. In Proceedings of the 18th ACM Workshop on Privacy in the Electroni Society (WPES'19). ACM, pp. 198-210.
42. *Ribeiro, M., Grolinger, K. and Capretz, M.A., 2016. MLaaS: Machine Learning as a Service. In Proceedings of the 14th International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 896-902.
43. Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), pp. 386-408.
44. Rowe, G. and Wright, G., 1999. The Delphi technique as a forecasting tool: issues and analysis. International journal of forecasting, 15(4), pp. 353-375.
45. Rumelhart, D.E., Hinton, G.E. and Williams, R.J., 1986. Learning representations by back-propagating errors. Nature, 323, pp. 533-536.
46. Schmidt, R. C., 1997. Managing Delphi surveys using nonparametric statistical techniques. decision Sciences, 28(3), pp. 763-774.
47. Schmidt, R., Lyytinen, K., Keil, M. and Cule, P., 2001. Identifying software project risks: An international Delphi study. Journal of management information systems, 17(4), pp. 5-36.
48. Sen, A., 2004. Metadata management: Past, present and future. Decision Support Systems 37(1), pp. 151-173.
49. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... and Hassabis, D., 2020. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), pp. 706-710.
50. Skulmoski, G. J., Hartman, F. T. and Krahn, J., 2007. The Delphi method for graduate research. Journal of Information Technology Education: Research, 6(1), pp. 1-21.
51. *Srimathi, H. and Krishnamoorthy, A., 2019. Personalization of Student Support Services using Chatbot. International Journal of Scientific & Technology Research 8(9), pp. 1744-1747.
52. *Stoyanovich, J., 2019. TransFAT: Translating fairness, accountability and transparency into data science practice. In Proceedings of the 1st International Workshop on Processing Information Ehtically (PIE) co-located with 31st International Conference on Advanced Information Systems Engeneering (CAiSE). CEUR-WS, pp. 1-10.
53. *Subbiah, U., Ramachandran, M. and Mahmood, Z., 2019. Software engineering approach to bug prediction models using machine learning as a service (MLaaS). In Proceedings of the 13th International Conference on Software Technologies (ICOSOFT). SCITEPRESS, pp. 879-887.
54. Sutton, R.S. and Barto, A.G., 1998. Introduction to reinforcement learning (Vol. 135). Cambridge: MIT Press.
55. *Tang, X.J. and Tay, Y.H., 2019. O-12: Xylorix: An AI-as-a-Service Platform for Wood Identification. In Program & Abstract of the IAWA-IUFRO International Symposium on Challenges and Opportunities for Updating Wood Identification. IAWA/IUFRO/CAF, pp. 31-32.
56. Turing, A. M., 1950. Computing machinery and intelligence. Mind, LIX(236), pp. 433-460.
57. von Briel, F., 2018. The future of omnichannel retail: A four-stage Delphi study. Technological Forecasting and Social Change, 132, pp. 217-229.
58. *Wang, L. and Lou, M., 2019. Complexity vs. performance: empirical analysis of machine learning as a service. In 2019 International Symposium on VLSI Design, Automation and Test (VLSI-DAT). IEEE, pp. 1-3.
59. Warth, J., Heiko, A. and Darkow, I. L., 2013. A dissent-based approach for multi-stakeholder scenario development—the future of electric drive vehicles. Technological Forecasting and Social Change, 80(4), pp. 566-583.

60. Wirth, R. and Hipp, J., 2000. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. Springer Verlag, pp. 29-39.
61. Wu, X., Zhu, X., Wu, G.-Q. and Ding, W., 2013. Data mining with big data. IEEE Transaction on Knowledge and Data Engeneering, 26(1), pp. 97-107.
62. Wynekoop, J. L. and Walz, D. B., 2000. Investigating traits of top performing software developers. Information Technology & People, 13(3), pp. 186-195.
63. *Yao, Y., Xiao, Z., Wang, B., Viswanath, B., Zheng, H. and Zhao, B., 2017. Complexity vs. performance: empirical analysis of machine learning as a service. In Proceedings of the 17th Internet Measurement Conference (IMC). ACM, pp. 384-397.
64. *Yousif, M., 2017. Intelligence in the Cloud – We Need a Lot of it. IEEE Cloud Computing, 4(6), pp. 4-6.
65. *Zhang, Y., Xu, F., Frise, E., Wu, S., Yu, B. and Xu, W., 2016. DataLab: a version data management and analytics system. In Proceedings of the 2nd International Workshop on BIG Data Software Engineering (BIGDSE'16). ACM, pp. 12-18.
66. *Zhao, S., Talasila, M., Jacobson, G., Borcea, C., Aftab, S.A. and Murray, J.F., 2019. Packaging and Sharing Machine Learning Models via the Acumos AI Open Platform. In Proceedings of the 17th Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 841-846.