CHALLENGES IN THE DEPLOYMENT AND OPERATION OF MACHINE LEARNING IN PRACTICE

Research paper

Baier, Lucas, Karlsruhe Institute of Technology, Karlsruhe, Germany, lucas.baier@kit.edu
Jöhren, Fabian, Karlsruhe Institute of Technology, Karlsruhe, Germany, fabian.joehren@web.de
Seebacher, Stefan, Karlsruhe Institute of Technology, Karlsruhe, Germany, stefan.seebacher@kit.edu

Abstract

Machine learning has recently emerged as a powerful technique to increase operational efficiency or to develop new value propositions. However, the translation of a prediction algorithm into an operationally usable machine learning model is a time-consuming and in various ways challenging task. In this work, we target to systematically elicit the challenges in deployment and operation to enable broader practical dissemination of machine learning applications. To this end, we first identify relevant challenges with a structured literature analysis. Subsequently, we conduct an interview study with machine learning practitioners across various industries, perform a qualitative content analysis, and identify challenges organized along three distinct categories as well as six overarching clusters. Eventually, results from both literature and interviews are evaluated with a comparative analysis. Key issues identified include automated strategies for data drift detection and handling, standardization of machine learning infrastructure, and appropriate communication and expectation management.

Keywords: Machine learning, Challenges, Interview study.

1 Introduction

Due to the large increase of data in recent years, various industries are trying to reap the benefits of this new resource for their service offerings. Machine learning (ML) is playing an important role in nearly all fields of business, ranging from marketing over governmental tasks to scientific-, health- and security- related applications (Chen et al., 2012). Furthermore, many companies rely on ML models deployed in their information systems for increasing the efficiency of their processes (Schüritz et al., 2016) or for offering new services and products. (Dinges et al., 2015). As Davenport (2006) describes, companies which are able to leverage their data sources through analytical tools achieve a substantial competitive advantage. But also for empirical science, ML enables novel ways of analyzing high-dimensional experimental data. This growth in popularity in both science and industry can also be explained by a massive increase in computational power (Jordan et al., 2015).
However, the wide-spread application of ML is rather young and therefore still confronted with many obstacles. Major challenges that have emerged recently in research and practice are datasets with high dimensionality (Cai et al., 2018), model scalability (Hazelwood et al., 2018), distributed computing (Z.-H. Zhou, 2017) and the live application of ML on streaming data (Z.-H. Zhou, 2017). In addition, related work has argued that published ML research is sometimes not driving sufficient real-world impact (Boutaba et al., 2018). The performance differences in developed algorithms rapidly diminish once applied onto real applications (Rudin et al., 2014). Furthermore, it has been criticized that research indeed performs evaluations on real-word datasets but that it does not appropriately communicate the results back to the

application domain (Wagstaff, 2012). This is closely related to the criticism on using unrealistic evaluation metrics (Rudin et al., 2014). These challenges mainly refer to technical issues. However, the successful implementation of a ML project also requires the consideration of organizational aspects. Therefore, in this work, we are interested in the predominant challenges during the practical implementations of ML projects. This leads to the following research question:

RQ: *Which challenges in the application and deployment of machine learning can we identify in practice?*

For answering this question, we perform at first a structured literature review of challenges named in literature which are organized along the categories pre-deployment, deployment and non-technical challenges. Subsequently, we conduct a study with 11 semi-structured interviews with ML practitioners working in various industries for identifying relevant challenges in their daily work. Subsequently, we perform a comparative analysis between challenges identified with the interviews in practice and the results from literature. In contrast to previous work focusing on technical challenges, we also identify non-technical ones such as a proper expectation management as well as challenges with creating new digital services based on ML. Our overview of challenges can guide the development of more realistic ML models in academia and can also be used as a support tool for practitioners in order to more efficiently plan and execute their ML projects.

The remainder of the paper is structured as follows: The next section covers related literature that we base our research on. In section 3, we describe our methodology for the interview study before we present our results in section 4. Section 5 discusses our results and compares literature and interview challenges. The final section describes theoretical and managerial implications, acknowledges limitations and outlines future research.

2 Related Work

Various challenges regarding the application of ML are considered in literature. In order to give a broad overview, we considered various domains and the whole life cycle of a ML project. We perform our literature review with AISel and Scopus as databases according to the methodology described in Webster et al. (2002). We draw upon CRISP-DM (Wirth et al., 2000) which is a standard process model for data

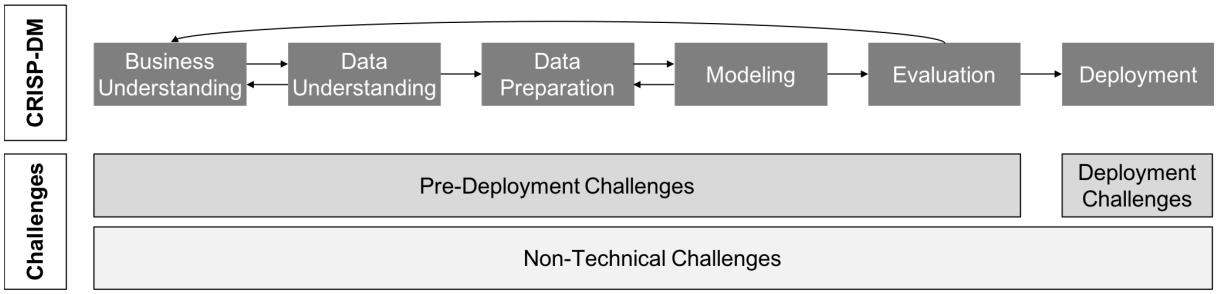

Figure 1: Conceptual phase model for categories of challenges

analytics projects for organizing the resulting challenges. However, some of the steps of CRISP-DM overlap and it is difficult to distinguish between them with regard to arising challenges in practice (e.g. data understanding and data preparation). Furthermore, business understanding is usually not covered in scientific literature. Therefore, we merged all challenges related to steps before the deployment of an ML model to pre-deployment challenges and the subsequent challenges to deployment challenges. Furthermore, we realize that ML projects often are also accompanied by other challenges which cannot be classified into the previous two categories. Therefore, the category non-technical challenges is added. Figure 1 introduces the applied categories.

Additionally, after performing the literature search, we merged the identified challenges into 6 distinct clusters consisting of similar challenges: Data structure, Implementation, Infrastructure, Governance,

Customer relation and Economic implications. Those are marked by blue boxes in Figure 2 which introduces the relevant challenges identified in literature.

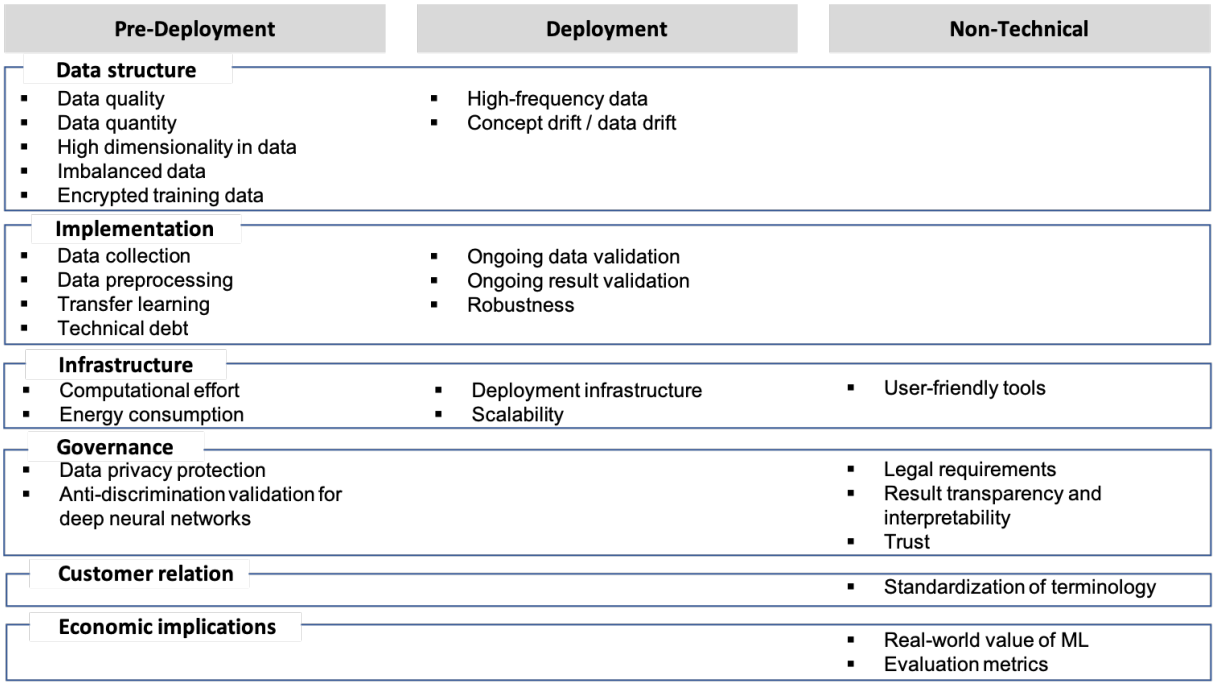

Figure 2: Challenges identified in literature

Pre-Deployment Challenges. Data is the fundamental basis of every ML project. The proper data structure with the right quality and also a sufficient amount of data samples is a prerequisite for a successful project. Often only a small amount of data being available is a problem (Boutaba et al., 2018; Brodley et al., 2012; Garcıa-Laencina et al., 2008; Zhang et al., 2018). If data is available, incomplete data, incorrect entries, or noisy features often make it difficult to achieve satisfying results (Baesens et al., 2014; Brodley et al., 2012; Kocheturov et al., 2018; Werts et al., 2000). One more problem is imbalanced or biased data. Even though a lot of solution approaches have already been presented, there is still room for improvement (Blenk et al., 2017; Kocheturov et al., 2018; J. Zhou et al., 2016). Applying ML on encrypted data sets is also challenging. However, even though some progress has been made for challenges such as training on small encrypted data sets, there is still a lot of work to do for training on large encrypted data sets (Graepel et al., 2012; Sarwate et al., 2013; Xie et al., 2014).

Moreover, big data brings along its own challenges. In many cases, it is still a huge problem to train algorithms on large amounts of data (Saidulu et al., 2017; Zhang et al., 2018). Especially, handling big data in reasonable time is challenging (Lopes et al., 2017). Furthermore, the high dimensionality of datasets in big data applications leads to a more complex feature engineering as well as requires other preprocessing steps (Cai et al., 2018; Domingos, 2012; Ferguson, 2017; Kocheturov et al., 2018; Sarwate et al., 2013). Furthermore, transfer learning for applying models on similar tasks across various domains (Suthaharan, 2014) can be challenging due to different data distributions. It is difficult to weight already learned patterns against information from new training data in this context (Jordan et al., 2015; Silver, 2011; Wang et al., 2016; Ying et al., 2015). Additional problems during the modeling phase are the concept of technical debt, describing the additional time needed in the future to adapt unclean code compared against clean code (Sculley et al., 2015). This is a challenges which arises during pre-deployment but which also might consequences during the deployment phase.

Infrastructure issues such as reducing the computational effort for model training and thereby lowering the memory requirements and energy consumption as well as increasing training and performance speed

are often mentioned challenges with regard to the model design (Jordan et al., 2015; Saidulu et al., 2017; Shafique et al., 2017; Xie et al., 2014; Zhang et al., 2018). These problems especially come along with big data sets. Solutions need to be found for making the models applicable in practice (Dietterich et al., 2008). In addition, data privacy protection and data security are governance challenges that need to be considered when applying ML models (Lopes et al., 2017). Legal frameworks like the European General Data Protection Regulation increase the complexity of the deployment of well-functioning solutions (Malle et al., 2017).
Since various authors use different terminologies when describing similar phenomena, we will use the terms "evaluation" and "validation" synonymously. Especially for 'black box' models like deep neural networks validation is challenging (Staples et al., 2016; Z.-H. Zhou, 2017). A certain level of transparency is necessary, if automated decisions are supposed to be based on such models in the future. It is crucial to ensure that the ML model does not discriminate based on any racial, sexual or other characteristic during its decision process (Anderson et al., 2015; Staples et al., 2016).

Deployment Challenges. During deployment, incoming data arriving with high frequency and large quantity can be challenging (Polyzotis et al., 2017). Concept drift, which describes a change in the distribution of input data or the distribution of the target variable, is an additional relevant challenge during deployment (Baier et al., 2019; Gama et al., 2014; Heit et al., 2016; Saidulu et al., 2017; Tsymbal, 2004; Widmer et al., 1996). Dietterich et al. (2008) name being able to handle changing distributions as one of the requirements for theoretical models to be applicable in practice.
Ongoing validation is an additional challenge for the implementation of ML models in practice. Algorithms developed and validated in research environments are not automatically applicable and easily validated for large data sets during deployment (Staples et al., 2016). This applies to validating incoming data with regard to quality and completeness as well as the resulting model predictions (Polyzotis et al., 2017). Furthermore, robustness is named as a major challenge. This refers, among others, to detecting and handling outliers appropriately. Moreover, reasonable results still need to be issued when the quality of the input data decreases (Boutaba et al., 2018; Hazelwood et al., 2018; Z.-H. Zhou, 2017). Ensuring robustness is described as very crucial and difficult, especially in autonomous driving applications (Koopman et al., 2017).
The scaling of small models developed on local hardware to deploying it on a large infrastructure with big amounts of data can cause problems. On the one hand, the infrastructure itself often leads to difficulties. Building up infrastructures with massive amounts of computing power (Hazelwood et al., 2018; Lopes et al., 2017; Shea et al., 2018), handling the energy consumption of those architectures (Hazelwood et al., 2018; Shafique et al., 2017) and working with infrastructures like mobile devices or cars (Koopman et al., 2017) is challenging. On the other hand, applying the algorithms on large amounts of data or on various types of infrastructures requires dedicated knowledge (Boutaba et al., 2018; Dyck, 2018; Parker, 2012). Ensuring that the models process incoming data and generate decisions within narrow time windows can be challenging, especially in use cases such as credit fraud detection (Baesens et al., 2014).

Non-Technical Challenges. The application of ML models for people with no background in data science is still quite challenging. Therefore, the introduction of user-friendly tools enabling non-technical employees to build their own models is required, since this would drastically increase the real-world impact of ML techniques (Dyck, 2018; Ferguson, 2017; Z.-H. Zhou, 2017). This is closely related to the concept of self-service analytics (Acito et al., 2014).
Legal requirements often pose a significant challenge for a machine learning project. This relates to data privacy protection as well as decisions on who is going to be accountable for false decisions based on ML models (Koopman et al., 2017). In addition, the results of ML models need to become more transparent and understandable for domain experts (Leung et al., 2016; Werts et al., 2000). Especially deep neural networks appear as a problem when it comes to transparency and interpretation (Nunes et al., 2017). However, in many domains results must be fully understandable to be really valuable (Nunes et al.,

2017; Rudin et al., 2014). A problem closely related to understanding and transparency is trust. Only if users really trust the results of ML models, they will rely on them when facing the challenge of making important decisions. Since the level of transparency for many ML model types is still low, also trust remains an open challenge (Baesens et al., 2014; Nunes et al., 2017; Shafique et al., 2017). Furthermore, the highly specialized and little standardized terminology used in ML is stated as a problem for novices in the field (Rudin et al., 2014; Wagstaff, 2012).
It is often argued that ML solutions developed in research have little or no real world value (Boutaba et al., 2018; Domingos, 2012; Sarwate et al., 2013; Werts et al., 2000). More realistic evaluations of model results need to be implemented to solve this problem (Heit et al., 2016; Rudin et al., 2014). Journals and editors need to support this development by requesting rigorous assessments of developed solutions under real world conditions (Wagstaff, 2012).
More standardization is needed for evaluating ML models (Shafique et al., 2017; Spangler et al., 2000) and the corresponding economic implications. On the other hand, evaluation metrics always have to be considered in the industry context where they are applied. Frequently, the same metrics with equal value ranges are compared for various application fields, even though the range implies completely different meanings (Wagstaff, 2012).
As shown, literature deals with a lot of different challenges for applied ML and motivation for research papers are often driven by real world problems. However, we want to examine if the relevant challenges in practice match with the ones stated in literature. The interviews intend to identify a potential gap between the challenges stated in literature and the ones named by practitioners.

3 Research methodology

In order to gain a comprehensive overview of the practical challenges of ML projects, 11 semi-structured expert interviews are conducted. Following the approach of Helfferich (2011), an interview guideline is used, structuring the interviews with regard to pre-deployment, deployment and non-technical issues.

3.1 Sampling

We apply a purposive sampling approach, including interview partners from various industries. Thereby, we aim to comprehensively cover occurring challenges of ML projects by including a variety of different perspectives and applications. Moreover, we ensure that different company sizes and maturity levels regarding data science projects are represented within the interview study. The interview partners (IP) are working in the following industries: Automotive (A) and other Manufacturing (M), Process (P), Power Generation (PG), Health Care (HC), Information Technology (IT), and ML as a Service (MLaaS). The consultants (CO) from the MLaaS companies cover the additional fields of Retail (R), Finance (F), E-Commerce (EC), Insurance (I), and Media (ME). Table 1 shows a complete overview of the different IP as well as the industry of their respective company.
All experts are developing ML solutions within specific projects in their daily tasks. Moreover, each of them has at least one year of experience, except two IP who have been working for six months in the field. Five of the eleven IPs work in ML consultancies and six of them hold ML positions within a specific company. Nine of the eleven experts live and work in Germany, one in the United States and one in Canada.

3.2 Data Collection and Analysis

All interviews are either conducted in person at the interviewee's office, via video call, or via phone call. The interviews were recorded after consent was granted and for further analysis transcribed. A qualitative content analysis (Krippendorff, 2004) is conducted to analyze the interviews. In order to remain open for the identification of new aspects and challenges of ML projects, the interview material is coded by

	Role		Industry									
	Consultant	Industry Expert	M/A	P	PG	HC	IT	R	F	EC	ME	I
IP α	X		X				X		X			
IP β	X							X		X		
IP γ		X					X					
IP δ		X	X	X	X							
IP ε		X	X									
IP ζ		X				X						
IP η	X								X			
IP θ		X	X	X								
IP ι		X					X					
IP κ	X		X							X		
IP λ	X		X					X		X	X	X

Table 1: Industry Overview of Interviewees

applying open coding. Two researchers independently conduct the analysis. The resulting code system is discussed and merged after each interview. As the involved researchers did not uncover additional insights after the fifth interview, the final coding system was fully developed, along with the corresponding coding rules. The remaining six interviews are used to evaluate the reliability of the coding system by applying the intercoder accordance. Therefore, the same two researchers code the remaining interviews, using the derived coding system. The number of matching codes per interview is computed, resulting in an average value across all interviews of 77,5%, which underscores the reliability of the derived coding system (Krippendorff, 2004). In order to be able to compare the results of the literature review (see Section 2) with the findings of the expert interviews, the derived codes were sorted according to the main categories 'non-technical challenges', 'pre-deployment', 'deployment'. A comparative analysis of ML challenges from literature with barriers of industry projects is provided in Section 5.

4 Results

This chapter introduces the challenges resulting from the interviews as well as first insights on best practices to deal with those challenges. Again, we differentiate between pre-deployment, deployment and non-technical challenges. Furthermore, we also apply the six clusters of challenges as defined in section 2. Figure 3 gives an aggregated overview of the identified challenges in both, the interviews (marked with an asterisk) as well as the literature (marked with a four corner star).
Usually, the deployment of ML models is performed by a specialized team of technical employees in collaboration with the corresponding department which is requesting a solution for its business problems. Therefore, in the following, departments requesting ML projects are referred to as 'customers'. We do not differentiate whether the service provider is an external ML consultancy or a dedicated ML team within the same company.

4.1 Pre-Deployment

Challenges referring to the data structure are frequently named by the interview partners, especially problems with data quality and quantity. Usually, data quality is examined before the start of a new project. However, a realistic assessment whether all required data sources are available is often only possible during the project (α, γ, ε). Furthermore, recognizing quality problems within the data is often rather difficult without domain expertise. Therefore, data scientists and domain experts need to collaborate closely to identify data quality problems. Imbalanced training data also complicates the application of

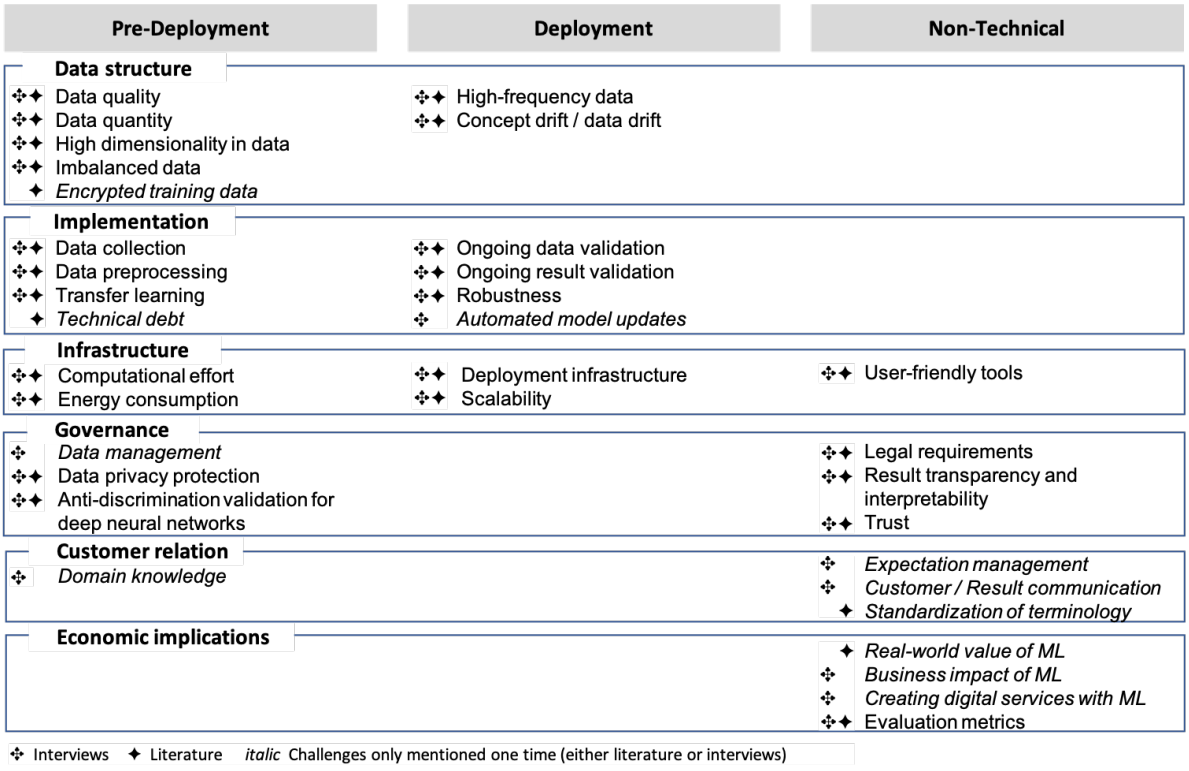

Figure 3: Challenges identified in interviews as well as in literature

ML models, e.g. in use cases to predict faulty products with a dataset with only very little faulty products at all (ε). Limited training data was also frequently mentioned as a challenge. However, some interview partners also stated that this does not affect their daily work. Especially when customers have a lot of experience with ML, they usually collect required data before the project start (λ).

Both problems, data with low quality and limited training data, are handled almost similar by all interview partners. First, domain knowledge is taken into account in order to increase data quality. Sometimes, this knowledge is also used as input for the model assumptions. Second, practitioners try to build initial ML model as simple as possible, which can provide reasonable results even with a small amount of data. In parallel, more data is collected and the implementation of the model continues as soon as new data is available. However, several interview partners also reported that projects are cancelled if data is too scarce (α,β,γ,ε,η). In health care, the problem of data collection is even more complex than in other domains (ζ). This is explained by two reasons: First, the progress of digitalization in general is slower in health care compared to other industries. Some data is not even available in digital form at all. Second, legal regulations for medical data are especially strict.

Data preprocessing is also named a fundamental problem (ε) because it is as a very time-consuming task which requires the vast majority of a project's time. Therefore, data preprocessing needs to be automated and accelerated (θ). Increasing the performance of ML algorithms and reducing their training time has no practical benefit, if data preprocessing remains as time-consuming as it is today (ι).

The actual modeling work is hardly mentioned as challenging. This refers to activities such as the decision on the type of algorithm or implementing the actual model which has been simplified enormously by the development of open-source frameworks. Only occasional challenges like transferring a built solution to another domain (θ) or use cases like autonomous driving (ε) are indicated. One interview partner (ι) explained that they are working on making deep neural networks more energy-efficient. Instead of using the whole range of available computing power, they want to develop an algorithm which only focuses on

the important part of a neural network to improve prediction performance.
Furthermore, governance issues such as legal and access rights often play an important role. Data management in companies across different industries is often organized poorly (γ, η). This is a challenge which is only mentioned during the interviews. Data access guidelines are usually very strict and complex. In general, several approvals are required to access relevant data (γ, ε, ζ). Before the actual work on data pipelines and modeling can start, it is often necessary to perform time-consuming tasks on data infrastructures (η). Hence, the wish for more sophisticated data structures in companies in general was stated (ε). Additionally, technical transparency is seen as a challenging, e.g. it is difficult to train deep learning algorithms which do not mimic the discriminatory behavior represented in the input data (β).

4.2 Deployment

Data with high volume as well as drifts in the input data are frequent challenges. Although there are several technical solutions for automatically recognizing shifts within the input data, like using Kafka input streams, manual checks are done most of the time (β, δ, κ). Changes within the input data are mentioned as a problem, especially for the validation of the model results (ε). Manual model adjustments are often performed to match the models to the new data distributions. Only in one case the models are able to adapt themselves automatically to data drifts (η).
In case of ML model updates, it is necessary to provide a neat documentation of all models including older versions. It needs to be documented which data has been used for training the model and under which conditions the model was performing well. In addition, an easy rollback to older versions must be available (β, δ). Therefore, a serving infrastructure with a proper model management is required. This allows an easy handling of different model versions as well as the opportunity to frequently update models (δ, ε). Furthermore, automated data pipelines pose a problem (κ) since they need to be able to combine database and ML model management. Further, templates for ML models and an automated, ongoing computation of prediction scores should be included. Cloud solutions offer standardized solutions for these challenges. Interview partners report fewer issues when using cloud services (λ). However, access to those is often restricted due to data privacy reasons or other restrictions. In general, robustness and stability are seen as major problems in deployment (β, δ, ι). Models must still provide reasonable results when facing minor data changes or a reduction in data quality.
In addition, ongoing validation of deployed ML solutions is mentioned as a problem. It is described as a time consuming, unstructured, and unstandardized process (ε, η). A key solution in most cases is a dedicated monitoring approach, which is either done automatically, manually, or with a combination of both. Continuous evaluation is the most important principle (β) and a clear definition of the corresponding metrics is required. Three different tests are proposed: consistency checks for the input data, continuous monitoring of the model predictions, and the effect of model predictions on prior defined KPIs. Such a continuous monitoring approach is the basis for a stable system. Manual sanity checks are a widely used mechanism to discover discrepancies in different areas within the pipeline. Further, the results are regularly investigated by domain experts (β, γ, δ, ε). In addition, automated consistency checks are used to compare current with previous prediction results. If predefined thresholds are violated, notifications can be triggered. Often, traffic light systems are used as a visualization tool (δ).
Infrastructure is one of the main problems during deployment of ML models. Challenges are not only related to deployment infrastructures for running the ML models, but also to setting up relevant data infrastructures (ζ, η). Three frequent challenges occur with regard to model deployment architectures according to our interview partners: First, data scientists often need to work with the infrastructure already available on the customer side. Therefore, data scientists have to adapt their solutions to various different infrastructures. This problem arises since approval processes for investments in new infrastructure are very time-intensive and complicated. Furthermore, many customers are very inexperienced with ML solutions and do not know if the investment is worth it. Second, standardized architectures for local solutions are scarce. Even if customers are willing to build up new infrastructure, it is difficult to install a consistent

local infrastructure. However, cloud solutions already offer this standardization extensively. Third, the actual deployment environment of ML models differs significantly. It requires fundamentally different approaches for running a model on either a large cluster in a manufacturing plant, directly in a car or on a mobile device.
Scaling up a model to deployment architectures also brings along additional challenges such as code parallelization. Only few programming languages are easy to parallelize and most of them are limited to computations in the internal memory. Usually, this is solved by building reasonable data partitions. Other approaches rely on using methods that allow out of core calculations (β) or adding more hardware (α, γ). However, the latter can be complicated. Methods for dimensionality reduction such as PCA or auto-encoders are applied for reducing the high dimensionality of datasets (δ). Usually, batch use cases can be handled by frameworks like Apache Spark (α, β). Real-time use cases which require immediate feedback, are more challenging and can require more advanced architectures (β). According to interview κ, only few use cases exist so far where big data infrastructures are really required. Many customers have so little experience that a well equipped, local server combined with a proper feature engineering approach is sufficient and significantly easier to run.

4.3 Non-Technical Challenges

During our interview study, we recognize that problems in the daily work with ML models are often also related to non-technical topics. Communication with the customer or translating ML results into real business impact are just two examples. Additionally, more standardization and user-friendliness for application of ML models are mentioned as a challenge. Easily applicable tools need to be developed in order to enable non-technical employees to apply ML models (δ).
The effect of ML models on business impact refers to two aspects: First of all, there are very different experience levels with regard to ML on the customer side. Many customers do not have a clear understanding of ML techniques and the corresponding benefits (α). This fact needs to be considered during the development of the customer relation. Second, providing transparency with regard to model results is challenging. It is difficult to convince customers to trust the ML results and to apply them for making crucial decisions (γ). Eventually, this challenge might be solved with technical solutions such as advanced frameworks for the visualization of important features. However, transparency itself remains a non-technical challenge. Standard ML metrics further complicate the problem of transparency. Most metrics are not easily understood by people without a ML background. Customers usually have difficulties in translating those metrics (e.g. accuracy) into relevant KPIs such as revenue (β).
Therefore, it is necessary to define individual, customer specific metrics at the beginning of a project to evaluate the results (ε) and the economic implications of the model. Furthermore, simpler, more understandable models are applied compared to complex deep neural networks. Customers often behave rather conservatively and select more understandable models over better performing ones (η). Technical solutions for increasing transparency are rare (η). Only few support tools such as Lime for visualizing deep learning models (ε) or Starlack for making R algorithms (η) are available.
Still, customer questions often cannot be answered in a satisfying way. Therefore, support from higher management positions is frequently required to communicate results and apply those accordingly (γ).
Although many companies already apply ML successfully in support systems, it is still difficult to create valuable digital services based on ML solutions (κ). This might also be related to the different accuracy needs for different domains. Changing legal requirements, such as data protection regulations, can further complicate the successful economic application of ML projects (β, ε).

5 Discussion

The interview results confirm the majority of challenges which are mentioned in related literature (c.f. section 2). However, we also identified gaps between the challenges stated in literature and the ones

mentioned in our study.
Researchers seem to be aware of many problems that practitioners are confronted with during the development of ML models in practice. However, solving these issues appears to be demanding, which might be due to two reasons. First, there might exist dedicated tools but those are not used by the majority of practitioners during their daily work either because those tools are not available (e.g. too costly) or because their usage is difficult. Second, adequate solutions for these problems have not been developed yet. However, we did not specifically ask whether the first or the second reason are the main driver of the respective challenge and therefore cannot make any statement about this.
Yet, the identified challenges are restricting practitioners in their daily work and therefore hindering a more widespread use of ML in practice. However, we cannot guarantee that individual highly advanced technology companies do not already possess sophisticated tools for some of the challenges which we discuss in the following. With this discussion, we want to raise awareness for the need of standardized solutions, which are easily applicable within companies with differing ML maturity levels.

Pre-Deployment. Data in general is a major challenge during model development in practice as well as in literature. Data collection and preprocessing requires the majority of time during ML projects. Data is often widely spread across the information systems of a company, is unstructured, and in a bad quality. Transforming data to the proper format usually requires a lot of manual work. The interviews specifically referred to data management with complicated access rights as a substantial challenge. Researchers typically are not confronted with this issue because they work with predefined datasets. Additionally, knowledge of several people (e.g. domain experts) needs to be merged to properly understand the data and raise the quality level of the data in practice. Furthermore, several interview partners referred to projects which were discarded after project start because of poor data conditions. This clearly indicates the critical importance of an adequate data structure as well as an appropriate data processing. Major problems in this category can easily jeopardize an entire machine learning project.
General solutions for the automation of data preprocessing and data structuring were directly requested by several of the interviewed experts. The problem has also been mentioned for decades in literature (e.g. Spangler et al., 2000). There is a clear need for tools supporting the whole data pipeline. Such tools could also support the faster evolution of ML techniques across different industries and application fields.
The selection of suitable machine learning algorithms and their improvements was named as challenging in both interviews and literature. However, several experts also noted that researchers too often focus on improving algorithms by small percentage points on statistical metrics while at the same time loosing sight of the complexity in real application domain. This problem has been stated before in literature, but we want to emphasize that researchers should also proof the applicability of their work in real-world environments. In that sense, after performing the interviews, we regard the identification and selection of the single best machine learning model as less critical. With an adequate preprocessing and a parameter optimization, a prediction model will perform well enough to bring a ML project to a successful end.
Literature, in contrast to the interviews, also refers to the problem of technical debt. This is certainly a challenge in practice, however it is less critical compared to other challenges since this issue usually can be solved with a corresponding time investment. Encrypted training data is another challenge only mentioned in literature. This might have not been a challenge so far in practice because encrypted data require rather sophisticated ML approaches. Currently, due to the novelty of deploying ML solutions, many companies still address the easiest and most promising use cases.

Deployment. Deploying ML models often is still a challenging task. This is also reflected by the fact that solutions in practice are often highly individual and require a lot of manual work. There are almost no standardized solutions for machine learning infrastructures in many domains. Hence, an individual infrastructure has to be built for many projects which is severely complicating the deployment. Unsuitable or missing infrastructure is a significant challenge for any ML project. If no proper deployment infrastructure can be set up by the project team, the entire project is prone to failure. Otherwise, infrastructure

issues still will significantly extend the timeline of a project due to long-lasting investment decisions, especially in larger companies.
Ongoing validation and data drifts are common challenges for deployed models. However, little automated strategies are available for handling these problems during deployment. The validation of deployed models is done with manual sanity checks in most cases. These often require the combined knowledge of data scientists and domain experts, which makes it a time consuming and complicated task. Data drifts are automatically detected by some models though, but are usually handled by manual model adjustments.
Automated model updates and adaptations are therefore a challenge which requires further research. This will either simplify and fasten the retraining process or even lead to tools which completely handle concept drifts autonomously without any human intervention. It is critical to develop a proper strategy to ensure the long-term validity of a deployed ML model already during the initial development of the ML model. Otherwise, the ML project is likely to fail to meet the performance expectations over time and customers with less technical experience might be disappointed and therefore be less open to new ML projects.
Proper infrastructure as well as ongoing validation both increase the robustness of deployed ML models in general which is widely confirmed as challenge during the interview study. Model robustness can also be enhanced by the algorithm development itself. Therefore, new or adapted algorithms should not only exhibit an increase in performance metrics, but also a higher level of robustness when confronted with erroneous data.

Non-Technical. Many ML projects are also considerably restricted by non-technical challenges. Transparency is indispensable for successful ML solutions and is often specifically demanded by customers and a proper understanding of model results is a necessary prerequisite for trust in ML models. Only trusted results will be considered for evaluating important decisions. Even though literature has extensively argued for more transparency, little progress has been achieved. This is especially true for deep neural networks which so far are very little explainable. Practitioners often use advanced visualization tools to increase transparency.
Creating more real-world value of ML solutions is an important challenge according to literature. Results from the interviews clearly support this statement. However, research papers usually are rather vague what real-world value actually means. During our interviews, we discovered that this challenges can be viewed from 4 different aspects:
First, it is necessary to express ML model results in terms of real-world business value and not in statistical metrics. In practice, evaluation metrics are usually defined individually for every ML projects and those metrics translate prediction results into important customer KPIs. However, this is very time-consuming and standardized real-world metrics could facilitate this process. Second, a proper expectation management with the customers during a ML project is crucial. Many customers are inexperienced in the field of ML which is why they cannot realistically assess what ML is able to accomplish. It is crucial that the project objectives are reasonably defined before any technical experiments start and that these objectives are also communicated appropriately to all stakeholders. These challenges have not been mentioned in literature so far, probably because researchers usually do not work in such complex project environments. However, a general framework for depicting the value and potential for economic applications of ML with corresponding business impact is in our opinion a valid research goal for resolving this challenge. Third, the communication with the customer as well as the comprehensible explanation of ML model results is important. Customers do not only want to understand the effect of ML results on their KPIs but they also want understand the influence of different features on the prediction results. In practice, this leads to the application of rather simple algorithms, even though more complex models usually easily outperform those. However, customers are often willing to accept lower performance in exchange for higher transparency. Fourth, creating valuable digital services or products based on ML model results is still quite challenging. It is difficult to convince people to pay for newly created services which are entirely based on ML or for existing services that are enhanced with ML capabilities.
Expectation management, an adequate customer communication and the creation of valuable digital

services based on ML are all critical challenges with regard to a more widespread use of ML. Concerning a single project, they will typically only lead to significant delays. However, if relevant stakeholders such as responsible line managers are not convinced by the capabilities of ML after the end of a project, this might have long-lasting consequences. Those managers are usually the ones who are identifying and providing use cases suitable for a ML application. Furthermore, they normally also provide the necessary funding for such a project. This means that if those stakeholders are not satisfied with the results after the execution of a project, success in future projects will be less likely.
Many of the introduced challenges actually go beyond the actual deployment of ML model solutions. However, CRISP-DM (Figure 1) as a standard process model ends with the evaluation of the overall project after deployment. Therefore, after having analyzed this large set of challenges, we argue for an extension of CRISP-DM because many activities such as an appropriate transfer of ML results or the ongoing monitoring and adaptation due to data drifts is not considered so far.

6 Conclusion and Outlook

The application of machine learning (ML) has spurred many new technological developments in both research and industry over the past years. However, many questions with regard to the application of ML in real-world applications are still unanswered. In this work, we identify typical challenges that are hindering ML practitioners in their daily work. We conduct a structured literature review as well as semi-structured interviews with 11 ML practitioners working in different industries.
Compared to publications addressing ML in a scientific context, our results show that practitioners do not only face traditional challenges such as data quality and data preprocessing, but are also confronted with a variety of additional problems during the deployment of ML solutions. This especially refers to a proper setup of the necessary infrastructure as well as solution strategies for handling concept drift and ensuring long-term validity of ML models. We therefore argue for more research with respect to these challenges since they can easily jeopardize the success of an entire project. Furthermore, practitioners frequently encounter non-technical issues such as the expectation management of customers (e.g. managers or non-technical employees) with regard to the deployed ML solutions as well as a proper communication of the results. Frameworks for depicting the value of ML can be a valuable resource in that case and could therefore be a valid research contribution.
Our research generates several contributions to the field of ML. First, we provide an overview of challenges of ML projects based on a structured literature review. These challenges are organized along the categories pre-deployment, deployment and non-technical challenges. Furthermore, we identified 6 overarching clusters of challenges: Data structure, Implementation, Infrastructure, Governance, Customer relation, Economic implications. Second, we provide an overview over challenges that ML practitioners are confronted with during their daily work (c.f. Figure 3). Based on both overviews and literature, we perform a comparative analysis, thereby, identifying similarities and differences between the challenges mentioned in literature, originating from a scientific context, and the practical barriers, which were identified through the interview analysis. These results have both implications for academia and industry. One the one hand, the total overview of identified challenges may be used to develop more realistic ML models in academia and provides guidance for future research. On the other hand, it serves as guidance for practitioners in the implementation of ML models.
Besides these contributions, our work faces a set of limitations. First, we conducted a limited amount of eleven interviews with ML practitioners. Furthermore, due to the availability of suitable interview partners, we could not cover all industry sectors with several interview partners. Nevertheless, we are confident about the completeness and validity of our results, as we did not encounter new challenges with the inclusion of new interview cases. Second, most of our interview partners are currently working in Germany, which might lead to a certain bias in our results. Third, due to the chosen qualitative approach, only limited statements can be made about the severity of one challenge in comparison to another, as well as about the prioritization of the different research needs.

Future work could overcome these shortcomings by performing an interview study with an international sample and could also identify corresponding best practices. During our study, we also realized that there are large differences in the perception of ML between ML experts and involved company managers. Therefore, a subsequent interview study with ML experts as well as with company managers would surely generate valuable insights. Additionally, based on these results, a larger quantitative study (e.g. survey-based) could be initiated and performed. This would allow the derivation of quantitative findings and the identification of the magnitude and severity of each challenge as well as the corresponding research need. Those findings could subsequently be used to derive holistic research priorities which promote the general progress of the field as a whole. In general, ML as a field has rapidly evolved over the past years. Therefore, it is necessary to continuously align the challenges occurring in the practical application with research pursued in academia.

Related literature has noted before that large parts of ML research are too narrowly focused on optimizing performance on benchmark datasets while not creating sufficient real-word value (Rudin et al., 2014). With our work, we want to initiate discussions and projects with the aim of closing the gap between academic and practical application. Solutions for the identified research needs can help to strengthen the practical implications of ML solutions.

References

Acito, F. and V. Khatri (2014). "Business analytics: Why now and what next?" *Business Horizons* 57 (5), 565–570.

Anderson, M. and S. L. Anderson (2015). "Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm." *Industrial Robot: An International Journal* 42 (4), 324–331.

Baesens, B., R. Bapna, J. R. Marsden, J. Vanthienen, and J. L. Zhao (2014). "Transformational issues of big data and analytics in networked business." *MIS quarterly* 38 (2), 629–631.

Baier, L., N. Kühl, and G. Satzger (2019). "How to Cope with Change?-Preserving Validity of Predictive Services over Time." In: *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Blenk, A., P. Kalmbach, W. Kellerer, and S. Schmid (2017). "o'zapft is: Tap Your Network Algorithm's Big Data!" In: *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*. ACM, pp. 19–24.

Boutaba, R., M. A. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O. M. Caicedo (2018). "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities." *Journal of Internet Services and Applications* 9 (1), 16.

Brodley, C. E., U. Rebbapragada, K. Small, and B. Wallace (2012). "Challenges and opportunities in applied machine learning." *AI Magazine* 33 (1), 11–24.

Cai, J., J. Luo, S. Wang, and S. Yang (2018). "Feature selection in machine learning: A new perspective." *Neurocomputing* 300, 70–79.

Chen, H., R. Chiang, and V. Storey (2012). "Business Intelligence and Analytics: From Big Data to Big Impact." *Mis Quarterly* 36 (4), 1165–1188.

Davenport, T. H. (2006). "Competing on analytics." *Harvard business review* 84 (1), 98–107.

Dietterich, T. G., P. Domingos, L. Getoor, S. Muggleton, and P. Tadepalli (2008). "Structured machine learning: the next ten years." *Machine Learning* 73 (1), 3–23.

Dinges, V., F. Urmetzer, V. Martinez, M. Zaki, and A. Neely (2015). "The future of servitization: Technologies that will make a difference." *Cambridge Service Alliance Executive Briefing Paper*.

Domingos, P. (2012). "A few useful things to know about machine learning." *Communications of the ACM* 55 (10), 78–87.

Dyck, J. (2018). "Machine learning for engineering." In: *Proceedings of the 23rd Asia and South Pacific Design Automation Conference*. IEEE Press, pp. 422–427.

Ferguson, A. L. (2017). "Machine learning and data science in soft materials engineering." *Journal of Physics: Condensed Matter* 30 (4).

Gama, J., I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia (2014). "A survey on concept drift adaptation." *ACM computing surveys (CSUR)* 46 (4), 1–37.

Garcıa-Laencina, P. J., A. R. Figueiras-Vidal, and J.-L. Sancho-Gómez (2008). "Machine learning techniques for solving classification problems with missing input data." In: *Proceedings of the 12th World Multi-Conference on Systems, Cybernetics and Informatics*.

Graepel, T., K. Lauter, and M. Naehrig (2012). "ML confidential: Machine learning on encrypted data." *International Conference on Information Security and Cryptology*, 1–21.

Hazelwood, K., S. Bird, D. Brooks, S. Chintala, U. Diril, D. Dzhulgakov, M. Fawzy, B. Jia, Y. Jia, A. Kalro, et al. (2018). "Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective." *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 620–629.

Heit, J., J. Liu, and M. Shah (2016). "An architecture for the deployment of statistical models for the big data era." *2016 IEEE International Conference on Big Data (Big Data)*, 1377–1384.

Helfferich, C. (2011). *Die Qualität qualitativer Daten*. Springer.

Jordan, M. I. and T. M. Mitchell (2015). "Machine learning: Trends, perspectives, and prospects." *Science* 349 (6245), 255–260.

Kocheturov, A., P. M. Pardalos, and A. Karakitsiou (2018). "Massive datasets and machine learning for computational biomedicine: trends and challenges." *Annals of Operations Research*, 1–30.

Koopman, P. and M. Wagner (2017). "Autonomous vehicle safety: An interdisciplinary challenge." *IEEE Intelligent Transportation Systems Magazine* 9 (1), 90–96.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Leung, M. K., A. Delong, B. Alipanahi, and B. J. Frey (2016). "Machine learning in genomic medicine: a review of computational problems and data sets." *Proceedings of the IEEE* 104 (1), 176–197.

Lopes, N. and B. Ribeiro (2017). "Novel Trends in Scaling Up Machine Learning Algorithms." *Machine Learning and Applications (ICMLA)*, 632–636.

Malle, B., P. Kieseberg, and A. Holzinger (2017). "Do not disturb? Classifier Behavior on Perturbed Datasets." *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 155–173.

Nunes, I. and D. Jannach (2017). "A systematic review and taxonomy of explanations in decision support and recommender systems." *User Modeling and User-Adapted Interaction* 27 (3-5), 393–444.

Parker, C. (2012). "Unexpected challenges in large scale machine learning." *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, 1–6.

Polyzotis, N., S. Roy, S. E. Whang, and M. Zinkevich (2017). "Data management challenges in production machine learning." In: *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, pp. 1723–1726.

Rudin, C. and K. L. Wagstaff (2014). "Machine learning for science and society." *Machine Learning* 95 (1), 1–9.

Saidulu, D. and R. Sasikala (2017). "Machine Learning and Statistical Approaches for Big Data: Issues, Challenges and Research Directions." *International Journal of Applied Engineering Research* 12 (21), 11691–11699.

Sarwate, A. D. and K. Chaudhuri (2013). "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data." *IEEE signal processing magazine* 30 (5), 86–94.

Schüritz, R. and G. Satzger (2016). "Patterns of data-infused business model innovation." In: *IEEE 18th Conference on Business Informatics (CBI)*. IEEE.

Sculley, D., G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison (2015). "Hidden technical debt in machine learning systems." *Advances in neural information processing systems*, 2503–2511.

Shafique, M., R. Hafiz, M. U. Javed, S. Abbas, L. Sekanina, Z. Vasicek, and V. Mrazek (2017). "Adaptive and energy-efficient architectures for machine learning: Challenges, opportunities, and research roadmap." *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 627–632.

Shea, C., A. Page, and T. Mohsenin (2018). "SCALENet: A SCalable Low power AccELerator for Real-time Embedded Deep Neural Networks." In: *Proceedings of the 2018 on Great Lakes Symposium on VLSI*. ACM, pp. 129–134.

Silver, D. L. (2011). "Machine lifelong learning: challenges and benefits for artificial general intelligence." *International Conference on Artificial General Intelligence*, 370–375.

Spangler, W. E., H. M. Chung, and F. C. Gey (2000). "Data Mining: A Brief Introduction to the Field and Research Community." In: *AMCIS 2000 Proceedings*.

Staples, M., L. Zhu, and J. Grundy (2016). "Continuous validation for data analytics systems." *Proceedings of the 38th International Conference on Software Engineering Companion*, 769–772.

Suthaharan, S. (2014). "Big data classification: Problems and challenges in network intrusion prediction with machine learning." *ACM SIGMETRICS Performance Evaluation Review* 41 (4), 70–73.

Tsymbal, A. (2004). *The problem of concept drift: definitions and related work*. Tech. rep. Computer Science Department, Trinity College Dublin.

Wagstaff, K. L. (2012). "Machine learning that matters." In: *Proceedings of the 29th International Conference on Machine Learning*, pp. 1851–1856.

Wang, W., M. Zhang, G. Chen, H. Jagadish, B. C. Ooi, and K.-L. Tan (2016). "Database meets deep learning: Challenges and opportunities." *ACM SIGMOD Record* 45 (2), 17–22.

Webster, J. and R. T. Watson (2002). "Analyzing the past to prepare for the future: Writing a literature review." *MIS quarterly*.

Werts, N. and M. Adya (2000). "Data Mining in Healthcare: Issues and a Research Agenda." In: *AMCIS 2000 Proceedings*.

Widmer, G. and M. Kubat (1996). "Learning in the presence of concept drift and hidden contexts." *Machine learning* 23 (1), 69–101.

Wirth, R. and J. Hipp (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*. Citeseer, pp. 29–39.

Xie, P., M. Bilenko, T. Finley, R. Gilad-Bachrach, K. Lauter, and M. Naehrig (2014). "Crypto-nets: Neural networks over encrypted data." *arXiv preprint arXiv:1412.6181*.

Ying, J. J.-C., B.-H. Lin, V. S. Tseng, and S.-Y. Hsieh (2015). "Transfer learning on high variety domains for activity recognition." In: *Proceedings of the ASE BigData & SocialInformatics 2015*. ACM.

Zhang, Y., J. Fu, C. Yang, and C. Xiao (2018). "A local expansion propagation algorithm for social link identification." *Knowledge and Information Systems*, 1–24.

Zhou, J., M. A. Khawaja, Z. Li, J. Sun, Y. Wang, and F. Chen (2016). "Making machine learning useable by revealing internal states update-a transparent approach." *International Journal of Computational Science and Engineering* 13 (4), 378–389.

Zhou, Z.-H. (2017). "Machine learning challenges and impact: an interview with Thomas Dietterich." *National Science Review* 5 (1), 54–58.