A Layered Quality Framework for Machine Learning-driven Data and Information Models

Shelernaz Azimi and Claus Pahl
Free University of Bozen - Bolzano, Bolzano, Italy

Keywords: Data Quality, Information Value, Machine Learning, Big Data, Data Quality Improvement, Data Analysis.

Abstract: Data quality is an important factor that determines the value of information in organisations. Data, when given meaning, results in information. This then creates financial value that can be monetised or provides value by supporting strategic and operational decision processes in organisations. In recent times, data is not directly accessed by the consumers, but is provided 'as-a-service'. Moreover, machine-learning techniques are now widely applied to data, helping to convert raw, monitored source data into valuable information. In this context, we introduce a framework that presents a range of quality factors for data and resulting machine-learning generated information models. Our specific aim is to link the quality of these machine-learned information models to the quality of the underlying source data. This takes into account the different types of machine learning information models as well as the value types that these model provide. We will look at this specifically in the context of numeric data, where we use an IoT application that exhibits a range of typical machine learning functions to validate our framework.

1 INTRODUCTION

Large volumes of continuously produced data are now omnipresent in many contexts. The Internet-of-Things (IoT) is a typical example where high volumes of a variety of data types are produced with high velocity (speed), often subject to veracity (uncertainty) concerns. We also refer to this type of data as big data (Saha and Srivastava, 2014).

In order to make sense out of this raw data originating from various sources, data needs to be structured and organised to provide information ready for consumption. More and more, Machine Learning (ML) is used to aggregate and derive non-obvious information from the data, thus enhancing the value of that information. Enhanced information can help to monetise data in the form of products or services provided. It can also aid an organisation in order to improve operational and strategic decision making. Machine learning makes this possible in a situation where manual processing and creation of functions on data is not possible due to time and space needs.

The problem in this context is, however, the impact of volume, variety, velocity and veracity of data on the quality and value of the information that is derived through a machine learning approach. Data quality plays a central role in creating value out of data (Heine et al., 2019). In order to conceptualise the problem, we need to extend a data quality framework to an ML function level (Azimi and Pahl, 2020b).

Thus, our contribution is a layered data architecture for data and ML function layers with associated quality aspects, consisting of

- a data quality model for raw data collected from different sources,
- a categorisation of machine learning functions,
- a quality model for machine-learning generated information models.

The problem of assessing quality of machine learning-generated information has been recognised (Ehrlinger et al., 2019), but a systematic categorisation is lacking. The novelty of the approach lies in, firstly, the layering of data and ML model quality based on dedicated ML function types and, secondly, when data quality might not be directly observable, we provide a new way of inferring quality.

This would enable to understand better the impact of data quality on enhanced information models for direct user access. Ultimately, this would also allow us to infer data quality deficiencies by only considering the machine-learned information models, thus allowing to remedy problems at data level and increase the overall value.

Azimi, S. and Pahl, C.
A Layered Quality Framework for Machine Learning-driven Data and Information Models.
DOI: 10.5220/0009472305790587
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 1*, pages 579-587
ISBN: 978-989-758-423-7

We focus here on numeric data that would for example be collected in technical or economic applications, neglecting text and image data here. We use IoT here as the application context in order to make qualities and impacting factors more concrete though an application context impacted by the big data factors volume, variety, velocity and veracity.

The remainder of the paper is structured as follows. In Section 2, we introduce essential background technologies. In Section 3, we review related work. Our quality framework is presented in Section 4. In Section 5, we motivate and validate the framework by referring to some use cases. Section 6 summarises the evaluation results, before we conclude in Section 7.

2 BACKGROUND

We look at the quality of information that is derived from data through a machine learning approach. Data quality and machine learning shall be introduced. Furthermore, we explain the role of IoT as a representative domain.

2.1 Data and Information Quality

Data represents a valuable asset in any enterprise context as a source for extracting information. Thus, quality is a critical requirement for any data communication and consumption, as be easily be motivated in IoT data services and their users. In IoT, things such as sensors produce volumes of data that are the basis in order to provide services for consumers. If data are inaccurate, extracted information and actions based on this will probably be unsound and erroneous. Sensors and other data collectors generally monitor a variable of interest (e.g., temperature, traffic, resource consumption, etc.) in the physical world (Azimi and Pahl, 2020a). The environments in which the collection of data takes place is often changeable and volatile in nature. Consequently, data is often uncertain, erroneous, noisy, distributed and voluminous. Some characteristics can depend on the context and the monitored phenomena, such as smoothness of variations, continuity and periodicity of production, correlation with other factors and statefulness (e.g., Markovian behavior).

Data quality refers to how well data meets the requirements of its consumers. The relevant aspects data quality are known as data quality dimensions (e.g., Accuracy, Timeliness, Precision, Completeness, Reliability and Error recovery). Based on this broader conceptualization, four main categories have been identified (Intrinsic, Contextual, Representational, Accessibility) based on 159 individual dimensions (Karkouch et al., 2016), see Table 1.

Data and information quality frameworks have been proposed (O'Brien et al., 2013), which we will review in the Related Work section. There is also a commonly accepted classification of (big) data aspects that can help in organising and managing the quality concerns (Saha and Srivastava, 2014): volume (scale, size), velocity (change rate/streaming/real-time), variety (form/format) and veracity (uncertainty, accuracy, applicability). All of these do, to a different extent, impact on our solutions.

2.2 Machine Learning

Machine learning (ML) is the study of algorithms and statistical models that computer systems use to perform a particular task, relying on patterns and inference and without being given precise instructions. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning tasks are classified into several broad categories. Supervised Learning, Unsupervised Learning and Reinforcement Learning are the most common ones (Mahdavinejad et al., 2018):

- In supervised learning, the algorithm builds a model from data that contains both inputs and desired outputs. Classification and regression algorithms are types of supervised learning. Classification is used when the output is a discrete number and regression when the output is continuous.
- In unsupervised learning, the algorithm builds a mathematical model from data that contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points.
- Reinforcement learning algorithms are given feedback in the form of positive or negative reinforcement (rewards) in a dynamic environment.

The quality of ML and the models that it produces through the different mechanisms has been an important concern for a long time ago (Cortes et al., 1995). Accuracy of the model in terms of the reality reflected for example is a central quality property.

2.3 Internet-of-Things

The Internet-of-Things (IoT) is the application framework that we consider here, where data that could be

Table 1: Categories of Data Quality Categories.

Data Quality Categories	Definition	Examples
Intrinsic	Categories that describe quality that is innate in or that inherently exists within data.	Accuracy, Reputation
Contextual	Categories describing the quality with respect to the context of tasks using data.	Timeliness, Completeness, Data volume
Representational	Categories describing how well data formats are representative and understandable.	Interpretability, Ease of understanding
Accessibility	Categories that describe how accessible (and in the same time secured) data are for data consumers.	Accessibility, Access security

high in volume, velocity and variety and is under veracity concerns (Saha and Srivastava, 2014) could be produced and processed by sensors and actuators.

We make here some assumptions. Firstly, that all data is numerical in nature, i.e., we do not consider text or image data and corresponding quality concerns regarding formatting and syntax here. We also assume that data can be stateful and also stateless. In that way, IoT is not the only application domain for our investigation and our results are transferable to other contexts such numeric data, but we focus on IoT here for motivation and evaluation (Li et al., 2018).

3 RELATED WORK

The related work shall be discussed in terms of three aspects: the data level, machine learning process perspective and machine learning model layer.

The data level was broadly discussed in (O'Brien et al., 2013) as a discussion paper. The data quality problems where classified into 2 groups of context-independent and context-dependant from data and user's perspective. These problems are both text and non-text based. Furthermore, in (Casado-Vara et al., 2018), a new architecture based on Blockchain technology was proposed as a concrete example in which an edge layer and a new algorithm were introduced to improve data quality and false data detection, which connected to our use of IoT as a sample context here.

Another similar work has been presented in (Sicari et al., 2016). In this paper, a lightweight and cross-domain prototype of a distributed architecture for IoT was presented, providing minimum data caching functionality and in-memory data processing. A number of supporting algorithms for the assessment of data quality and security were presented and discussed. In the presented system, users could request services on the basis of a publish/subscribe mechanism, with data from IoT devices being filtered according to user requirements in terms of security and quality. The prototype was validated in an experimental setting characterized by the usage of real-time open data feeds presenting different levels of reliability, quality and security.

The process perspective was presented in (Amershi et al., 2019). Nine stages were specified for an ML workflow in which some of the stages were data oriented. ML workflows are highly non-linear and contain several feedback loops which may loop back to any previous stage. This workflow can become even more complex if the system is integrative, containing multiple ML components which interact together.

Finally, the machine learning model layer has been studied in (Plewczynski et al., 2006), (Caruana and Niculescu-Mizil, 2006) and (Caruana and Niculescu-Mizil, 2005). Different algorithms and approaches were introduced and used in these papers which were mostly building on supervised learning algorithms. They observed that different methods can have different applications. They also examined the effect that calibrating the models via Platt scaling and isotonic regression has on their performances.

For concrete ML techniques, there are also specific quality metrics applied. For instance, (Kleiman and Page, 2019) discuss the area under the receiver operating characteristic curve (AUC) as an example for classification models. We respond to that by associating primary quality concerns to the identified model types here later on.

A different direction has been described in (Sridhar et al., 2018). In this paper, the authors motivate the need for, define the problem of, and propose a solution for model governance in production ML. They showed that through their approach, one can meaningfully track and understand the who, where, what, when, and how an ML prediction came to be.

The quality of data in a machine learning approach has been investigated in (Ehrlinger et al., 2019), where an application use case is presented. A systematic coverage of quality aspects, is however not attempted. However, to summarise, no joint discussion of quality concerns across the layers exist yet, which we aim to address here. (Ehrlinger et al., 2019)

is the most relevant paper to our work but, they did their research on just 1 use case and we intend to do a broader research.

4 DATA AND INFORMATION QUALITY FRAMEWORK

Information is the result of organising and structuring raw data coming from data-producing sources such as sensors in IoT environments.

4.1 Information Value

Information is based on data that can be put to a use. Firstly, information has direct financial value, i.e., the information can be monetised in the form of externally provided products or services. Secondly, the information can be used internally to support decision making at different levels, such as strategic or various forms of operational decisions. We can illustrate the value aspect in different application cases. We choose weather and mobility here as sample domains:

- Weather: paid weather forecasting services that aggregate and predict weather are common examples, i.e., direct monetization of the data and information takes place by the provider.
- Mobility: long-term strategic decisions, e.g., city planning, can be based on extracted and learned road mobility volume and patterns.
- Mobility: short-term operational planning, e.g., event management in a city or region can be based on common and extraordinary mobility behaviour derived from raw monitored data for past events.
- Mobility: immediate operation, e.g., self-adaptive traffic management systems such as situation-dependent traffic lights where the behaviour of adaptive traffic light controllers can be learned.

These examples demonstrate the different value aspects for specifically some use case domains.

4.2 Data and Information Quality Layers

Our central hypothesis is that information, as opposed to just data, is increasingly provided through functions created using a machine learning (ML) approach. IoT is a sample context were typically historical data is available that allows functions to be derived in the form of machine learning models.

We distinguish here three ML function types: predictor, estimator and adaptor, see Table 2. These reflect different usage context of ML techniques. The predictor addresses future events; the estimator deals with calculations or estimations irrespective of a state notion; and the adaptor calculates adjustments to a system in order to achieve a goal.

Based on this initial assumption, we present as the core of our framework a layered data architecture, see Figure 1, that captures qualities of the data and the information function layer.

- *Source Data Layer*: The base layer is the source layer consisting of raw, i.e., unstructured and unorganised data from IoT sources in our context. We can distinguish here context-dependent and context-independent quality properties. We follow here the frameworks presented in (O'Brien et al., 2013) and (Thatipamula, 2013), but adjust this to numerically-oriented data (i.e., exclude text-based and multimedia-based data sources).
  - context-independent data quality: e.g., missing/incomplete, duplicate, incorrect/inaccurate value, incorrect format, outdated, inconsistent/violation of generic constraint.
  - context-dependent data quality: e.g., the violation of domain constraints.

  Data quality needs to take into account data that are missing, incorrect or invalid in some way. In order to ensure data are trustworthy, it is important to understand the key dimensions of data quality to assess the cause of low quality data.
- *Information Model Layer*: The upper layer is an enhanced information model.
  - In order to define a quality framework for the information function, we considered as input for function quality the following structural model quality: completeness, correctness, consistency, accuracy and optimality that can be found in the literature (Plewczynski et al., 2006), (Caruana and Niculescu-Mizil, 2006).
  - Based on these we define a function quality notion for each of the function types[1], see Table 4. Note, that is essential here to assess the quality of the function provided by the ML models, which emerge in different types, such as predictors, estimators or adaptors.

In Figure 1, the source data is grouped into reality and rules aspects (intrinsic data quality category, see Table

[1] In the literature, also ethical model or function qualities such as fairness, sustainability or privacy-preservation can be found. Since there is some uncertainty about their definition, we will exclude these here.

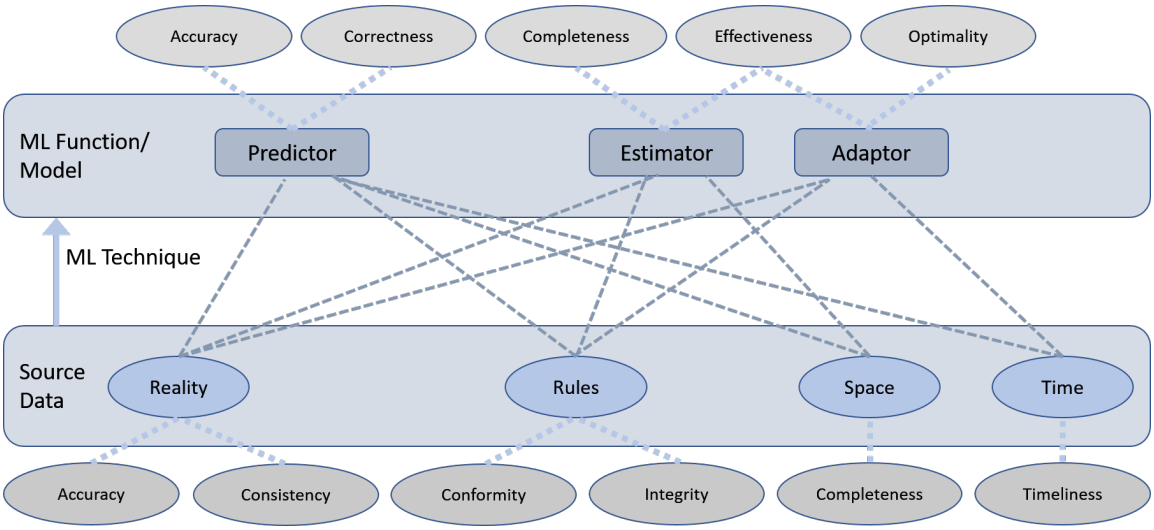

Figure 1: Layered Data Architecture.

Table 2: ML Function Types.

Function Type	Definition
predictor	a predictor predicts a future event in a state-based context were historical data is available.
estimator	an estimator or calculator is a function that aims to calculate a value for a given question, which is an estimation rather than a calculation if accuracy cannot be guaranteed.
adaptor	an adaptor is a function that calculates setting or configuration values in a state-based context where a system is present that can be reconfigured to produce different data.

Table 3: Data Quality Definitions.

Quality	Aspect	Quality Definition
Completeness	Space	Completeness is defined as expected comprehensiveness. Data can be complete even if optional data is missing. As long as the data meets the expectations then the data is considered complete.
Consistency	Reality	Consistency is the degree to which labeller annotations agree with one another. In other word, how often the labeling was correct.
Conformity	Rules	Conformity means the data is following the set of standard data definitions like data type, size and format.
Accuracy	Reality	Accuracy is the degree to which data correctly reflects the real world object or an event being described, i.e., how close a label is to the real world.
Integrity	Rules	Integrity means validity of data across the relationships and ensures that all data in a database can be traced and connected to other data.
Timeliness	Time	Timeliness references whether information is available when it is expected and needed. The timeliness depends on user expectations.

1) and space and time aspects (contextual data quality category), into which we organised the six individual qualities. At the ML model layer, the three function types predictor, estimator and adaptor are shown, which each of them having their primary quality concern attached. Across the layer, we have shown cross-layer dependencies. Here we can see a mesh of dependencies, with only adaptors not strictly requiring space qualities (i.e., allow systems to work in the case of incompleteness by not taking an action) and estimators not essentially based on a state/time notion.

Machine learning connects the two layers by mapping data into information models. We expect the mapping by the ML approach to have some properties. Critical here is conformance, i.e., the resulting functions must accurately represent the lower data layer. In some situations, we need to refine the quality classification. For the adaptor function, effectiveness

Table 4: ML Model Quality Definitions.

Quality	Function	Quality Definition
Correctness	Predictor	Correctness is a Boolean value that indicates whether a prediction was successful
Accuracy	Predictor	Accuracy is the degree to which a prediction was successful
Completeness	Estimator	Completeness is the degree to which a estimator covers the whole input space
Effectiveness	Estimator / Adaptor	Effectiveness is a Boolean value that indicates the correctness of a calculation
Optimality	Adaptor	Optimality is a Boolean value indicating whether the optimal solution has been reached

and optimality are criteria that often involved multiple goals. For example, for the primary goal 'effective' for one aspect (which could be a performance threshold in a technical system), we could have as secondary goal 'optimality' for another aspect (which could be energy or amount of resources sent to maintain performance). In order to see how these function qualities are calculated, see Table 5.

Table 5: ML Functions, Qualities and ML Techniques.

Function	Sample Quality	ML technique
predictor	accuracy correctness	regression classification
estimator	effectiveness completeness	clustering clustering (if there is no cluster)
adaptor	effective optimal (e.g. minimal)	classification regression, reinforcement

We can relate ML function quality to ML techniques. We look at the different function types individually. In practical terms, the complexity of the quality calculation is of importance, since in an implementation, the ML function assessment would need to be automated. Here, (i) complexity is a concern and (ii) there is also need to wait for actual observable result event (adaptor) as an example.

5 ILLUSTRATED USE CASE

We already illustrated the information value aspect for the Mobility and Weather domains, indicating that ML functions provide value for monetization through services/products and for decision support for strategic (long-term), operational (mid-term/short-term) and adaptive (short-term/immediate) needs.

We now investigate the Mobility case further, which will actually also involve weather data in order to combine different varieties of data. This serves here as an illustration, but helps also with the validation of the concepts we introduced in our framework. This is based on a real case study we are working on with a provider of ICT services for public administration in a regional context. The raw data sets from the sensor sources are of the two different domains:

Table 6: Data Quality Observations for Mobility Use Case.

Quality	Observation
incomplete	can arise as a consequence of problems with sensor connectivity and late arrival of data (causing incompleteness until the arrival)
integrity (duplicate)	sensors might be sending data twice (e.g., if there is no acknowledgement)
inaccurate	as a consequence of sensor faults
integrity (format failure)	if temperature data is sent in Fahrenheit instead of Celsius as expected
timely (outdated)	if either the observed object has changed since data collected (road capacity has changed) or data that has arrived late
inconsistent	where generic consistency constraints such as 'not null' in data records are violated

- *road traffic data*: number of vehicles (categorised) [every hour, accumulated]
- *meteorological data*: temperature and precipitation [every 5 minutes]

From this, a joined data set emerges that links traffic data with the meteorological data. Since we cannot assume the weather and traffic data collection points to be co-located, for each traffic data collection point, we associate the nearest weather data collection point.

The first component of our framework is raw data and its associated qualities. Quality of the raw data can be a problem, as shown in the following cases presented in Table 6.

Machine learning shall be utilised to create ML information models and derive different types of information through the ML function types :

- *prediction*: predicted number of vehicles for the next 5 days at a certain location,
- *prediction*: predicted level of traffic (in 4 categories light, moderate, high, very high) for the next 5 days at a certain location
- *estimation*: estimation of average number of vehicles in a particular period (which needs to be abstracted from concrete weather-dependent numbers in the data).
- *estimation*: estimation of the type of the vehicle such as car or motorbike.
- *adaptation*: determination of suitable speed limits to control (reduce) accidents or emissions.

The ML model creation process can use different techniques, including decision trees, random forests, KNN, neural networks etc. This is largely driven by a need for accuracy. In practice, a model will be created for each traffic location. For example, a neural network can be used to create a model for traffic level predication. ML model creation (training) takes into account historical data, which in our case is a full year of meteorological and traffic data for all locations.

The purpose is to create information models to support the following objectives across several value types presented in Table 7.

Table 7: Value Types for the Mobility Use Case.

Value type	Objective	ML function
strategic	road construction	prediction/ estimation
operational	holiday management	prediction
adaptive	speed limits	adaptation

Four functions across three function types shall now be described in more detail in terms of functionality and quality. The quality analysis of these functions is presented in Table 8.

What this information model and function analysis shows is that in our mobility application we cover the following cases: (i) all information value types (strategic, operational, adaptive) are covered, (ii) all ML function types (predictor, estimator, adaptor) are covered, (iii) all ML function qualities are relevant and applicable. Thus, this serves as a demonstration of the suitability of the quality framework we presented for common data collection, processing and analysis applications with numeric data. In addition to this case study, we evaluate central properties of our solution further in general terms below.

6 DISCUSSION

The evaluation aims at validating the proposed quality framework. Partly, the traffic use case in the previous section serves as a proof-of-concept validation of the concepts. However, here we cover the criteria more systematically and comprehensively. The evaluation criteria for our quality framework are:

- completeness of the selected qualities at both data and ML model levels,
- necessity of all selected qualities: required for the chosen use case domains,
- transferability across different domains beyond IoT and specifically mobility.

6.1 Observations

The above criteria shall be addressed individually.

6.1.1 Completeness and Necessity

In order to achieve a high degree of completeness, we consulted the literature on data quality and ML model quality and took respective frameworks on board. Some additional ones that are less structural in nature exist. These are sometimes referred to a ethical properties such as fairness, privacy-preserving, sustainable and address personal or societal concerns (Rajkomar et al., 2018). As there is no agreed definition for them in the community, we left them out here.

We report on transferability between domains below. There, we observed that all quality factors were applicable at least once, thus confirming the overall necessity of all factors.

The use case above has demonstrated completeness and necessity for the given application case, as we remarked at the end of that section.

6.1.2 Transferability

In order to address transferability, we considered data sets from a range of selected domains:

- weather: a range of different meteorological attributes,
- buildings: looking at an air-condition and heating system as a self-adaptive system that reacts to temperature and usage.
- people: has been considered as part of the buildings domain,

Table 8: Use Case ML Function Analysis.

Value	Function Type	Function	Construction	Quality
strategic	estimator	The long-term strategic aspect is based on traffic, but not weather. The estimated average number of vehicles over different periods is here relevant.	supervised learning – classification	*effective*: allows useful interpretation, i.e., effective road planning *complete*: available for all stations
operational	predictor	The operational aspect needs to predict based on past weather and past traffic, taking into account a future event (holiday period here). Concrete predictions are traffic level and traffic volume (number of cars)	supervised learning – classification	*correct*: right traffic level is predicted *accurate*: number of cars predicted is reasonable close to the later real value
	predictor	A second operational function could determine the type of car, e.g., if trucks or buses should be treated differently	unsupervised learning – clustering	*correct*: right vehicle is determined *accurate*: categories determined are correct for correct input data
adaptive	adaptor	a self-adaptive function that changes speed limit settings autonomously, guided by an objective (such as reducing accidents or lowering emissions).	unsupervised learning – reinforcement learning	*effective*: speed reduction is effective. *optimal*: achieves overall objects with the proposed action

- technical system: as an autonomous system, cloud computing environments have been evaluated as self-adaptive resource management systems that rely to some degree on prediction.

We found most of the concepts of our quality framework applicable in all domains, but all relevant in at least one domain. This overall confirms the transferability of the concepts. Due to space limitations, the results are not presented here.

6.2 Threats to Validity

Machine learning comprises various learning mechanisms suitable for processing numeric data, text and other multi-media formats such as images. In order to make qualified statements about data and information qualities, we restricted ourselves to numeric data with the respective qualities and ML processing types. Not all data quality concerns from the literature were included here. For instance, fairness, sustainability or privacy-preservation are relevant from a more societal than technical perspective. However, since there is no agreement in the community and these are still under investigation, these were excluded.

We choose IoT as the application domain, which might be too specific, but we aligned the discussion with the 4V model of big data (volume, variety, velocity, veracity) showing that these characteristics apply. Some frameworks also add 'Value' as a fifth concern to the big data characteristics (Nguyen, 2018). We also discussed this value aspect.

The use case discussion covers the selected ML function types, but also considers the different ML techniques such as supervised and unsupervised learning with classification, regression, clustering and reinforcement learning being applied.

7 CONCLUSIONS

Raw data is without additional processing of little value. More and more, machine learning can help with this processing. A critical observation in this context is the need to address quality across the two layers that emerge – the raw source data and the ML-supported information models.

We presented a quality framework that combines quality aspects of the raw source data as well as the quality of the machine-learned models derived from the data, We provided a fine-granular model covering a range of quality concerns organised around some common types of machine learning function types.

Machine learning is still expanding into different application areas. More complex information and knowledge models can be expected in the future. Digital twins is such an advanced concept that refers

to a digital replica of physical assets such as processes, locations, systems and devices, in which ML-generated models based on measured and monitored data from the real world form the basis for further analysis. These are often based on IoT-generated data with enhances models and function provided through machine learning. We plan to investigate deeper the complexity of these digital twins and the respective quality concerns that would apply.

As other future work, our ultimate goal is to close the loop mapping functional problems back to their origins by identifying the symptoms of low quality precisely and map these to the root causes of these deficiencies.

REFERENCES

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *International Conference on Software Engineering (ICSE 2019) - Software Engineering in Practice track*. IEEE Computer Society.

Azimi, S. and Pahl, C. (2020a). Particle swarm optimization for performance management in multi-cluster iot edge architectures. In *International Conference on Cloud Computing and Services Science CLOSER*. SciTePress.

Azimi, S. and Pahl, C. (2020b). Root cause analysis and remediation for quality and value improvement in machine learning driven information models. In *22nd International Conference on Enterprise Information Systems - ICEIS 2020*. SciTePress.

Caruana, R. and Niculescu-Mizil, A. (2005). An empirical comparison of supervised learning algorithms using different performance metrics.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 161–168, New York, NY, USA. Association for Computing Machinery.

Casado-Vara, R., de la Prieta, F., Prieto, J., and Corchado, J. M. (2018). Blockchain framework for iot data quality via edge computing. In *Proceedings of the 1st Workshop on Blockchain-Enabled Networked Sensor Systems*, BlockSys'18, page 19–24, New York, NY, USA. Association for Computing Machinery.

Cortes, C., Jackel, L. D., and Chiang, W.-P. (1995). Limits on learning machine accuracy imposed by data quality. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, pages 239–246. MIT Press.

Ehrlinger, L., Haunschmid, V., Palazzini, D., and Lettner, C. (2019). A daql to monitor data quality in machine learning applications. In Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A. M., and Khalil, I., editors, *Database and Expert Systems Applications*, pages 227–237, Cham. Springer International Publishing.

Heine, F., Kleiner, C., and Oelsner, T. (2019). Automated detection and monitoring of advanced data quality rules. In Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A. M., and Khalil, I., editors, *Database and Expert Systems Applications*, pages 238–247, Cham. Springer International Publishing.

Karkouch, A., Mousannif, H., Moatassime, H. A., and Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73:57 – 81.

Kleiman, R. and Page, D. (2019). Aucμ: A performance metric for multi-class machine learning models. In *International Conference on Machine Learning*, pages 3439–3447.

Li, H., Ota, K., and Dong, M. (2018). Learning iot in edge: Deep learning for the internet of things with edge computing. *IEEE Network*, 32:96–101.

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., and Sheth, A. P. (2018). Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks*, 4(3):161 – 175.

Nguyen, T. L. (2018). A framework for five big v's of big data and organizational culture in firms. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5411–5413. IEEE.

O'Brien, T., Helfert, M., and Sukumar, A. (2013). The value of good data- a quality perspective a framework and discussion. In *ICEIS 2013 - Proceedings of the 15th International Conference on Enterprise Information Systems*, volume 2.

Plewczynski, D., Spieser, S. A. H., and Koch, U. (2006). Assessing different classification methods for virtual screening. *Journal of Chemical Information and Modeling*, 46(3):1098–1106. PMID: 16711730.

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., and Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872.

Saha, B. and Srivastava, D. (2014). Data quality: The other face of big data. In *2014 IEEE 30th International Conference on Data Engineering*, pages 1294–1297. IEEE.

Sicari, S., Rizzardi, A., Miorandi, D., Cappiello, C., and Coen-Porisini, A. (2016). A secure and quality-aware prototypical architecture for the internet of things. *Information Systems*, 58:43 – 55.

Sridhar, V., Subramanian, S., Arteaga, D., Sundararaman, S., Roselli, D. S., and Talagala, N. (2018). Model governance: Reducing the anarchy of production ml. In *USENIX Annual Technical Conference*.

Thatipamula, S. (2013). Data done right: 6 dimensions of data quality. https://smartbridge.com/data-done-right-6-dimensions-of-data-qua\-lity/. Accessed on 2020-01-16.