CHALMERS
UNIVERSITY OF TECHNOLOGY

Identifying and managing data quality requirements: a design science study in the field of automated driving

Citation for the original published paper (version of record):
Pradhan, S., Heyn, H., Knauss, E. (2024). Identifying and managing data quality requirements: a design science study in the field of
automated driving. Software Quality Journal, 32(2): 313-360.
http://dx.doi.org/10.1007/s11219-023-09622-8

Software Quality Journal (2024) 32:313–360
https://doi.org/10.1007/s11219-023-09622-8

Identifying and managing data quality requirements: a design science study in the field of automated driving

Shameer Kumar Pradhan[1] · **Hans-Martin Heyn**[1] · **Eric Knauss**[1]

Accepted: 21 February 2023 / Published online: 11 May 2023

Abstract
Good data quality is crucial for any data-driven system's effective and safe operation. For critical safety systems, the significance of data quality is even higher since incorrect or low-quality data may cause fatal faults. However, there are challenges in identifying and managing data quality. In particular, there is no accepted process to define and continuously test data quality concerning what is necessary for operating the system. This lack is problematic because even safety-critical systems become increasingly dependent on data. Here, we propose a Candidate Framework for Data Quality Assessment and Maintenance (CaFDaQAM) to systematically manage data quality and related requirements based on design science research. The framework is constructed based on an advanced driver assistance system (ADAS) case study. The study is based on empirical data from a literature review, focus groups, and design workshops. The proposed framework consists of four components: a *Data Quality Workflow*, a *List of Data Quality Challenges*, a *List of Data Quality Attributes*, and *Solution Candidates*. Together, the components act as tools for data quality assessment and maintenance. The candidate framework and its components were validated in a focus group.

Keywords Advanced driver assistance systems · Data quality · Data quality attributes · Data quality challenges · Data quality workflow · Requirements engineering

1 Introduction

Successful deep learning requires a large volume of data during the design and operation of such systems (Rusk, 2016). Data used for training and operation is crucial in achieving the desired behavior of a deep learning system (Sun et al., 2017). Consequently, there is

✉ Shameer Kumar Pradhan
shameer.pradhan@uhasselt.be

Hans-Martin Heyn
heyn@chalmers.se

Eric Knauss
eric.knauss@cse.gu.se

[1] Department of Computer Science and Engineering, University of Gothenburg and Chalmers, Göteborg 40530, Sweden

a need to identify data quality challenges and systematically define relevant data quality attributes. However, there needs to be a systematic procedure to determine and manage data quality. Today, most of the data quality assessment information for the deep learning system is based on undocumented expert knowledge, especially during pre-processing of input data (Holstein et al., 2019).

An advanced driver assistance system (ADAS) is designed to make driving comfortable and safe by enabling drivers to make the right decisions (Ziebiński et al., 2017). The system assists in overtaking other vehicles, parking, and detecting obstacles. ADAS can also execute emergency braking and lane change independently. These systems are inherently safety-critical because they can intervene by braking and steering the vehicle. To enable all these functions, ADAS employs a perception system, which deploys deep learning and encounters a large volume of data during the design and operation phase (Fayyad et al., 2020). Such systems for ADAS include traffic sign recognition and road obstacle detection. Because of functional safety decomposition, the perception system will inherit functional safety requirements from ADAS. In turn, the deployed deep learning models in the perception system will also have to comply with functional safety requirements. Consequently, this means that the data used for training and testing the deep learning models must not compromise the safe function of the deep learning model.

This study aims to understand data quality requirements in the context of safety-critical systems like ADAS. Divergence from the expected system behavior can mean the difference between a safe journey and a fatal accident. The behavior of machine learning in general, and deep learning in particular, depends on the data, especially the quality of the data provided for training, validation, and inference at runtime. A lack of quality data might compromise the decision-making capabilities of the driver in the context of automated driving, which can result in a fatal accident. Thus, the data used for training the system should be appropriate for successfully operating in a real-world implementation. Similarly, data used for validation should be appropriate for determining whether the system will work as intended. Finally, during runtime, the inference must be based on data with a quality that resembles training and validation data quality; otherwise, it will be impossible to guarantee that the system is working within certain boundaries. Providing unsuitable data, i.e., data of poor quality, will lead to undesired system behavior and impact efficiency (Madnick et al., 2014; Challa et al., 2020).

1.1 Research questions and objectives

We formulate two research questions to guide our study:

Research question 1 (RQ1): What are the relevant data quality challenges in deep learning systems?
Research question 2 (RQ2): What constitutes a requirements framework for data quality management in deep learning systems?

Answering the first research question helps identify data quality challenges. Identification of such challenges can, in turn, help in devising solutions for those challenges. The second research question helps develop a series of components for a candidate framework whose goal is to help researchers, practitioners, and other stakeholders identify the data quality challenges, understand data quality attributes, and manage data quality overall.

The objectives of this study are as follows:

- To identify challenges associated with data quality for deep learning systems such as that can be found in ADAS;
- To understand data quality requirements for such systems;
- To devise a set of solutions for identifying and mitigating data quality challenges.

The primary contributions of this study are the identification of relevant data quality challenges and the development of a series of artifact components that assist in the identification and reduction/mitigation of such challenges. By understanding the identified data quality challenges, we establish a candidate framework that could lead to a framework that supports stakeholders in identifying and maintaining data quality and requirements towards data. According to McMeekin et al. (2020), a methodological framework "provides structured practical guidance or a tool that supports its user through a process in a step-wise manner."

We position the candidate framework devised in this paper as a stepping stone towards a comprehensive framework for understanding the data quality challenges and attributes for data-driven developments such as deep learning in ADAS.

The scope of the study is limited to establishing a candidate framework for data quality in the training and testing of deep learning models and, thus, does not relate to concrete data types produced by individual sensors. We study data quality requirements by exploring data quality challenges and attributes. The data collected for this study originates mainly from the past experiences of the experts. A candidate framework comprising various components is proposed based on data collected via interviews, focus groups, surveys, and literature review.

The remainder of this article is structured in the following manner. Background and related work are presented in Section 2. Similarly, Section 3 provides the study's methodology and design using automated driving as a case study. Section 4 provides the result of the study in the form of a candidate framework, including a set of primary components and their evaluation. The resulting candidate framework and its implication to researchers and practitioners are discussed in Section 5. Finally, Section 6 concludes the article and provides potential future directions for this study.

2 Background and related work

With the rise of distributed systems, data soon became a key concern. Standards such as ISO 25010 on software and data quality can guide the handling of data quality aspects for software systems ISO (2011). However, the standard was drafted before the rise of machine learning in the late 2010s. It aims to guide software architecture decisions instead of data selection in data-centric applications (Haoues et al., 2017).

A data quality framework for distributed computing environment by Fletcher (1998) proposes a measure called *Data Quality Risk Exposure Level (DQREL)*. DQREL is an attribute-dimensions matrix with eight data dimensions and three data attributes. As stated by the author, the DQREL matrix can be used to understand "data quality pitfalls" in a system.

A first step towards identifying data quality requirements is understanding the expectations for the final ML systems. Sandkuhl (2019) studied the expectations of two projects—one in financial industries and the other in ML and data science. The author devised a method component to understand the organizational context of ML, which can be used to conduct ML requirements analysis and, finally, analysis of data availability based on the elicited requirements towards the ML system.

That requirements towards the ML system directly result in requirements towards data quality, has been shown by Sessions and Valtorta (2006). The authors show that data quality impacts the effectiveness of machine learning algorithms. They devise procedures for developing robust and practical algorithms using data quality assessments. They evaluate the need for good data quality by developing and testing three Bayesian networks. However, assessing and managing the data quality of large datasets is a challenging task, as shown by Cai and Zhu (2015). The challenges of data quality they identify include difficulty in data integration, a large volume of data, fast-changing data, and a need for more data quality standards and frameworks. The authors propose a dynamic assessment process for data quality to identify these challenges. Another framework for data quality assessment and monitoring was developed by Batini et al. (2007). Based on the Basel II operational risk evaluation methods, the authors devised a data quality assessment methodology called *ORME-DQ*, which contains four phases for data quality risk prioritization, identification, measurement, and monitoring. The authors develop an architectural framework composed of five modules that support the phases of the assessment methodology.

The importance of such data quality assessment methods has also been shown by Fujii et al. (2020). The authors devised a set of guidelines for the quality assurance of AI. These guidelines connect data quality, model robustness, system quality, process agility, and customer expectation. They evaluated their proposed guidelines through a survey, with over 77% of the participants agreeing on their usefulness.

Among the five challenges in requirement engineering for ML-based applications identified by Vogelsang and Borg (2019), the elicitation of data required is one of them. The authors identified a gap between the tools used by data scientists to control data quality and requirement engineering connecting data quality requirements to customer expectations.

The Open Measured Data Management Working Group has developed a vendor-neutral platform called OpenMDM[1] to manage measured data. Automotive companies primarily use this platform to build in-house applications. It can, however, also be used to develop other solutions. It includes components and concepts that can be used to "compose applications for measured data management systems." OpenMDM can manage measurement data, evaluation results, and descriptions.

Other data management frameworks, such as datasheets for datasets proposed by Gebru et al. (2021), do not explicitly connect data quality attributes to data requirements. The "dataset nutrition label framework" introduced by Holland et al. (2020) provides an extendible approach for data scientists to compare different datasets summarized as labels. However, the framework requires a list of relevant data quality attributes and needs to explain how data quality challenges can be solved.

We propose the contribution of this study as a blueprint toward a framework for identifying and managing data quality attributes. Unlike previous studies that mainly investigated individual aspects of data quality, this study provides a consolidated tool that includes data quality challenges, related attributes, and solution candidates to overcome the data quality challenges. The main difference from previous approaches is that this proposed framework is extendable to data quality challenges, attributes, and solutions. Based on a case study, this article will provide many examples of data quality challenges, attributes, and solutions entered into the proposed framework.

[1] https://openmdm.org

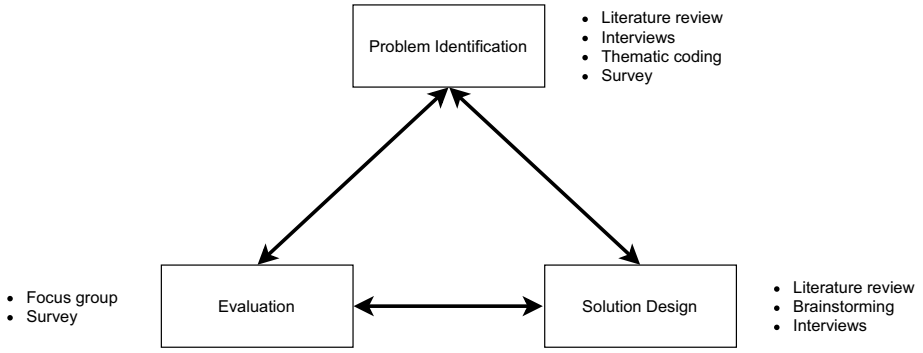

Fig. 1 Stages of design science research

3 Research method

Design science research (DSR) was performed for this study. According to Hevner et al. (2004), DSR is a problem-solving process enabled by developing and evaluating novel artifacts as solutions to problems. The DSR methodology is applicable in various domains, including software, human-computer interaction, and system design.

The study was performed in three cycles, with each cycle focusing primarily on one of the stages of DSR, namely, problem identification, solution design, and evaluation. However, other tasks were also updated if new information or idea was generated, irrespective of the stage. Data quality challenges were identified during the problem identification stage with the help of a literature review and expert interviews. The framework was devised during the second cycle, the solution design stage. The identified challenges and the framework were evaluated in the third cycle, the evaluation stage. All stages are illustrated in Fig. 1.

3.1 A case study in data quality for automated driving

Automated driving is adopted as the case study for this research. In this research, we conduct a case study to evaluate data quality challenges in automated driving. The study was conducted in collaboration with a Swedish Tier 1 supplier of automotive systems for original equipment manufacturers (OEMs), which designs, manufactures, and sells software and hardware systems for occupant protection, ADAS, collaborative, and automated driving. These systems include vision, radar, lidar, thermal sensing, electronic controls, and human–machine interfaces. We argue that this company is a representative for developing systems for automated driving, as it has customer relations with several OEMs worldwide, and it is one of the largest Tier 1 suppliers for perception systems used in automated driving in Europe.

Sampling strategy We employed a mixture of convenience sampling (Sedgwick, 2013) and purposeful sampling (Suri, 2011) techniques during the selection of the experts. The industry partner supported the selection of experts for this study and provided us with the experts based on our requirements regarding their expertise and area of work. We asked

the company to provide us with experts with a wide variety of experiences and positions involved in the development of automated driving functions to obtain a broader perspective and receive more diverse feedback on our interview questions (Palinkas et al., 2015). Our main selection criterion was the active involvement in product development for ADAS functions that use some form of machine learning.

3.1.1 First cycle: problem identification

During problem identification, one investigates the research objective from different perspectives in sufficient detail to support the design of a solution (Peffers et al., 2007). While it makes sense to focus on problem identification in the first cycle, understanding the problem should be revisited iteratively even during the other cycles of the DSR (Knauss, 2021). Similarly, although the focus was on solution design during the second DSR cycle, problem identification and evaluation were also consciously considered. Feedback from the evaluation stage was also used to further refine the problem understanding and solution design.

The first cycle involved interviews and a literature review as the primary source for identifying *data quality challenges*. The interviews, which were recorded and transcribed, were conducted via Microsoft Teams, an online communication tool. The data quality challenges were segregated using data-driven thematic analysis.

Based on the previously formulated research question, we developed an interview guide as Farooq and de Villiers (2017) state that a well-developed interview guide helps devise a better structure for the interviews. Furthermore, feedback received from interviews can be helpful in further refining and rephrasing the interview questions. Based on the outcome of previous interviews, questions were tuned accordingly to fill the knowledge gap for other interviews.

The goal of the interviews in the first cycle was to identify *data quality challenges*. Interviewees A–E, listed chronologically in Table 1, are the interviewees during the first cycle. Interviewees F–H in the same table are the interviewees during the second cycle. Five interviewees are experts from the case company, two additional interviewees are experts from two partner companies of our case company, and one expert is a research partner of the case company within an EU Horizon 2020 research project. We chose to add additional experts outside the case company to check the validity and transferability of the answers we received from within the case company.

Table 1 List of interviewees

ID	Role	Team	Experience
A	Research specialist	Research	25 years
B	Functional safety engineer	Driver assistance systems	5 years
C	Feature tech lead	Vision pre-development	22 years
D	Group manager	ADAS platform development	17 years*
E	Technical lead	AI and ML	15 years
F	Development manager for road traffic management	Traffic management	25 years
G	Product owner	Ground truth	6 years
H	Engineering technical fellow	Research and innovation	23 years

*9 years experience in automotive

The interviews were transcribed and thematically coded. Data-driven coding was used in the thematic analysis of the interviews of the first cycle, as described by Gibbs (2007). In such a technique, codes are based on the words used in the interviews.

A survey was conducted to understand the appropriate severity of the identified challenges. Interview participants from the first cycle and additional participants from a requirements engineering workgroup of a deep learning research project associated with the case company[2] participated in the survey.

While preparing the survey questionnaire, the identified challenges were divided into five categories. For each category, the survey participants were asked to rank the challenges by the level of severity. They were asked to rate the categories as well. The modified scale ranged from 1 to 6, with 1 being the least severe challenge and 6 being the most formidable challenge. A scale with an even number of alternatives was deliberately selected to induce the participants to "pick a side," as suggested by Cox (1980).

An algorithm to calculate a metric called *Challenge Score* was developed. The algorithm uses the ranking of individual challenges in their respective categories and the Likert scale value given to those categories to calculate a *Challenge Score*. The value is normalized over the total number of challenges in the respective category and the number of survey participants. More details about the algorithm and associated formula can be found in the accompanying data package.[3]

3.1.2 Second cycle: solution design

After identifying the problem in the first DSR cycle, the primary focus of the second DSR cycle was on solution design. The artifact was designed to meet the stakeholder requirements and resolve the identified challenges by building on the early prototypes from the first cycle.

A series of artifact components, which collectively form the Candidate Framework for Data Quality Assessment and Maintenance (CaFDaQAM), was designed as part of the solution design step. The components are explained in Sect. 4 of this article. Results from a literature review, the first round of interviews, the first survey, and the group brainstorming sessions between the researchers were used to devise the components and their content. We also conducted additional interviews in this cycle to verify the developed components. Furthermore, some of the questions asked during the interviews were open-ended to encourage brainstorming between the researchers and the interviewees.

The interviews of the second cycle were also thematically coded and analyzed. Unlike the thematic coding of the interviews of the first cycle, descriptive coding and analytic coding techniques were used to thematically code the interviews of the second cycle (Gibbs, 2007), (Skjott Linneberg & Korsgaard, 2019) because we were focusing on verifying the findings of the first cycle.

We used four deductive codes in this study. Those were *confirmation of a pre-identified challenge*, *confirmation of a proposed solution*, *rejection of a pre-identified challenge*, and *rejection of the proposed solution*.

[2] The group consists of participants of Work Package 2 of the *Very Efficient Deep Learning in the IoT* (VEDLIoT) research project in which the case company is actively involved. The research project aims to apply the proposed data quality framework for its use cases in distributed deep learning for automotive systems, home automation, and industrial IoT. See www.vedliot.eu for more details.

[3] https://doi.org/10.7910/DVN/Y6ORUV

3.1.3 Third cycle: evaluation

The third cycle of this study focuses primarily on the evaluation of the candidate framework. A preliminary evaluation was already conducted as part of the study's first and second cycles. For example, the interviewees were presented with the artifact components and solutions in a preliminary design phase during the second cycle. The presentation was done to gather their feedback regarding those components and solutions.

The evaluation was primarily done using a focus group and a survey. A focus group session was conducted to validate the candidate framework components. The focus group participants included researchers and engineers from academia and industry with experience in automated driving development, deep learning, and data quality. The session was conducted for 2 h with five participants: two from academia and three from the industry. The participants were confronted with questions to brainstorm regarding the association between the challenges, the data quality attributes, and the candidate framework components. They also shared their ideas and thoughts through discussion.

Finally, a comprehensive survey questionnaire was sent to members of the VEDLIoT requirements engineering workgroup. Ten participants submitted a response. However, the participants' identities could not be determined as the survey did not ask for their names to maintain anonymity. This survey aimed to validate the components of the candidate framework. It asked the participants to provide a Boolean response to the appropriateness of individual fields for the templates of the candidate framework components. In the same way, questions regarding data quality challenges, their association with data quality attributes, and their effect on deep learning models were asked in the survey.

3.1.4 Calculation of challenge score

During the first iteration, 27 data quality challenges were identified through interviews and a literature review. A way to rank the challenges was necessary for the effective analysis. *Challenge Score* ranks the identified challenges in terms of their severity, i.e., whether a challenge is more pressing or less.

The computation of the *Challenge Score* is based on the response from the survey conducted to rank the challenges. The survey contained two types of questions; one type of question asked the participants to provide a value of significance based on a Likert scale to five sets of challenges, and another type of question asked to rank individual challenges inside the five sets of challenges.

As there are two types of responses to two types of questions, their results need to be combined. The *Challenge Score* combines both types of responses in one final value. For each respondent, the value they provide for the comprehensive sets of challenges is recorded. The highest-ranked challenge in a challenge set is given the highest numerical value. Decreasing numerical values are assigned to remaining challenges in the particular challenge set. E.g., if there are four challenges in a challenge set, the highest-ranked challenge is given a value of 4, the second highest-ranked is given a value of 3, and so on.

For each challenge, the assigned numerical value is multiplied by the value given by that particular participant for the challenge set of that particular challenge. This process was repeated for all of the participants and challenges. The product values calculated for all participants for individual challenges are summed. The final *Challenge Score* is calculated by dividing this sum by the number of challenges in the particular challenge set and by dividing the result by the total number of participants, which is done to normalize the final value.

4 Candidate Framework for Data Quality Assessment and Maintenance (CaFDaQAM)

This section presents the final artifact titled *Candidate Framework for Data Quality Assessment and Maintenance* (CaFDaQAM), which was developed during the study. It includes four components listed in Table 2, namely a *Data Quality Workflow*, a *List of Data Quality Challenges*, a *List of Data Quality Attributes*, and *Solution Candidates*. Components here mean a series of tools that can be used, independently or in combination, to identify and manage data quality requirements. This section will outline each of the components. Furthermore, for each of the components, more details, implementations, and literature references are provided in the artifact package of this article.[4] In the following, we define *attribute* as "a concept providing qualitative information about a specific object" (Statistical Office of the EU, 2020).

4.1 Component I: Data quality workflow

This component presents a step-by-step workflow for assessing and managing data quality and requirements. It includes six steps, as shown in Fig. 2. Most of the steps can be performed in parallel, as depicted by the dotted line in Fig. 2. Loops indicate that the steps can be done iteratively. The components of CaFDaQAM can be associated with the different steps of the workflow, as depicted in Table 2. The workflow was developed through brainstorming with experts. Furthermore, it was presented to the industry practitioners working with the case study during the focus group session to collect feedback for its evaluation.

S1 Identify data quality challenges In this step, challenges concerning data quality can be identified from several sources. Examples of primary sources of data collection are interviews, field studies, and surveys. Research papers and books can be used as secondhand sources as well. Furthermore, the collected challenges can be divided into different categories. In this study, they were categorized into five groups relating to *data availability*, *data management*, *data source*, *data structure*, and *data trust*.

S2 Collect and organize data quality attributes In this step, data quality attributes can be identified from various sources. E.g., sources such as research papers, proceedings papers, books, standards, technical reports, Internet articles, and interviews can be used to identify the attributes. Data quality attributes can also be elicited from interviews. A single attribute can also represent differently phrased data quality attributes. E.g., *understandability* and *ease of understanding attributes* can be represented by the same attribute.

S3 Associate data quality challenges and data quality attributes Data quality challenges and quality attributes can be associated with each other after their identification. The association means that a certain data quality challenge affects a certain data quality attribute. There is a many-to-many relationship between data quality challenge and data quality attribute, i.e., one challenge can affect more than one attribute, and one attribute can be affected by more than one challenge. For instance, *accuracy* (attribute) is affected by *data drop*, *incomplete data*, etc. (challenges); and *data drop* (challenge) can affect *accuracy*, *completeness*, etc. (attributes).

[4] https://doi.org/10.7910/DVN/Y6ORUV

Table 2 Candidate Framework components and their purpose (*refer to Fig. 2 for the related steps in the workflow component)

ID	Component	Purpose	Step*
I	Data Quality Workflow	Provides a step-by-step workflow to assess and manage data quality	
II	List of Data Quality Challenges	Provides a template for challenges	S1, S6
III	List of Data Quality Attributes	Provides a template for data quality attributes. Also includes information regarding which challenges affect a particular attribute, metrics, and their formula	S2, S3, S4, S6
IV	Solution Candidates	Provides a template for solution candidates to reduce or mitigate the identified challenges. Also includes requirements specifications and implementation details (flowcharts) of the solutions	S5, S6

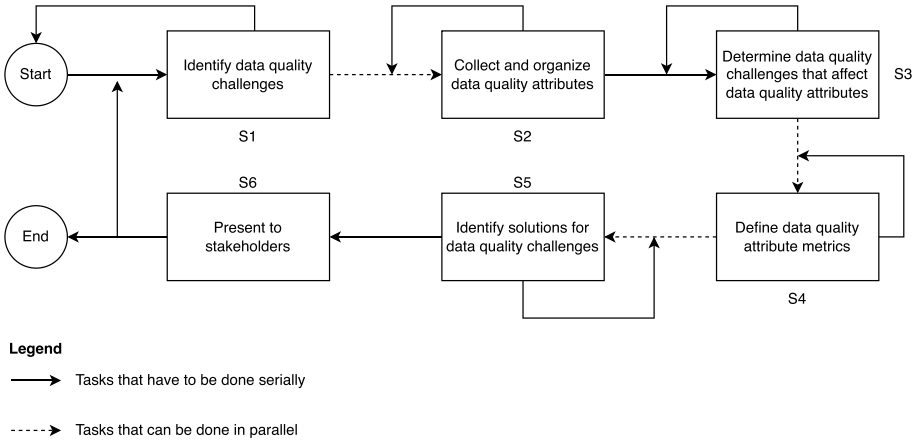

Fig. 2 *Data Quality Workflow* artifact component

However, there can be those data quality attributes that are not affected by any identified challenge and data quality challenges that do not affect any attribute.

S4 Define data quality attribute metrics Metrics to measure data quality attributes are formulated in this step. The metrics help to put a quantitative value on the attributes. E.g., *degree of accuracy* (metric) helps to measure *accuracy* (attribute). It gives a quantifiable value for the attribute. Furthermore, formulae can be devised to calculate the metrics. E.g., the *degree of accuracy* can be calculated as a ratio of the number of correctly labeled data records and the total number of data records. The formulae are mostly dependent on the context of the application.

S5 Identify solutions for data quality challenges A way of improving data quality attribute metrics, thus improving quality attributes, is to determine candidate solutions for the data quality challenges that affect the attributes. If the challenges can be mitigated or reduced, it will help improve the data quality attributes. For instance, finding a solution for *data drop* (challenge) and implementing it in the system process could result in lesser data being dropped, thus improving the *completeness* (attribute). Several sources, such as research papers, technical reports, and books, can identify solutions. Teams can also brainstorm and devise new solution candidates for the challenges. An effective way to validate solution candidates is to implement them as tests in part of a system.

S6 Present to stakeholders As the final step, identified data quality challenges, attributes, and solution candidates should be presented to appropriate stakeholders. They could be higher management, other colleagues, or customers. A suitable form of presentation should also be decided.

4.2 Component II: List of data quality challenges

Table 3 presents the template of *List of Data Quality Challenges* component. It includes eight fields validated by the participants of the focus group as well as the second survey. The participants were asked to decide whether a certain field was required or not for a particular

Table 3 Template for *List of Data Quality Challenges* artifact component

Field	Description	Example
Name	Name of the data quality challenge	Low labeled data volume
Reference	Reference that denotes the identification of the challenge	Interviews
Description	Description of the data quality challenge	In the training dataset, the volume of labeled data is significantly lesser than that of the unlabeled data. Since a large volume of data is unlabeled, the unlabeled data is useless, and the deep learning models need to be properly trained. E.g., if only 30% of the traffic signs in a scene are labeled, it would be "more difficult for the neural network to learn traffic signs since there are quite a lot of traffic signs among the negative samples."
Type	Type of challenge. Could be *Availability*, *Management*, *Source*, *Structure*, or *Trust*.	Availability
Directly affects AI Functions	Boolean value to denote whether the data quality challenge directly affects AI functions or not	Yes (1), No (0) (Note: All participants in the focus group affirmed that this challenge affects AI functions)
Challenge Score	A calculated value that denotes the ranking of the data quality challenge in terms of severity	Survey 1 - 4.333 (Rank 1/31), Survey 2 - 3.750 (Rank 1/25)
Responsible Stakeholders	People and departments in an organization responsible for handling the challenge	Data Collection Department, Data Manager, Data Collector
Impact Level	Degree to which the challenge affects the AI models. Could have values such as *High*, *Medium*, *Low*.	*High*

component. All participants responded that all fields except one (the source) apply to the component. The *source* field was agreed upon by 75% of the participants. The challenges are related to the case as they are identified by the experts from the case company. The challenges identified in the case study were entered into the template and are also provided in the artifact package.[5]

In response to the first research question (RQ1), in total, and at the end of the study, 27 data quality challenges were identified from elicitation methods such as interviews and literature review. During the course of the study, ten challenges were identified in our literature review analysis and interview data. Nine other challenges were only found in interview data, without a matching report in related work. The remaining eight challenges were identified only in the literature review. Figure 3 depicts the number of challenges retrieved from various sources, such as interviews and literature reviews, as well as the methods employed to validate the identified challenges. The challenges are divided into five broad categories: data availability, data management, data source, data structure, and data trust. We will list the identified challenges here; a complete description of all challenges, including more details on each challenge, is available in the artifact package[6] accompanying this article. As an extract from the artifact package, the challenges under the category *data availability challenges* are detailed in Appendix A.

Data availability challenges affect the data availability during processing by AI models. The challenges categorized under this challenge set are *Data Delay*[*], *Data Drop*[**], *Incomplete Data*[*], and *Low Labeled Data Volume*[**].

Data management challenges are related to data management and operations performed on them. The challenges categorized under this challenge set are *Data Acquisition*[***], *Data Ownership*[*], *Expensive Procedure*[**], *Imbalanced Dataset*[***], *Improrer Data Transfer*[*], *Large Volume of Data*[***], *Manual Data Collection*[***], *Manual Data Labeling*[**], *Redundant Data*[*], *Regulatory Compliance*[***], *Reliance on Suppliers to Raise Error*[**], and *Time Consuming*[**].

Data source challenges are those caused by and due to the source of the data. The challenges categorized under this challenge set are *Data Dependent on External Conditions*[**], *Lack of Variety in Test Environment*[**], *New Data Type*[*], and *Wrongly Calibrated / Defective Sensors*[*].

Data structure challenges are related to the format and structure of the data. The challenges categorized under this challenge set are *Fragmented Data*[***], *Incompatible Data Format*[***], *Outlier Data*[*], and *Unstructured Data*[***].

Data trust challenges are caused due to the lack of transparency in the data and its quality to extract meaningful information. The challenges categorized under this challenge set are *Incorrect Labeling*[*], *Lack of Good Data from Simulations*[**], and *Noise*[*].[7]

4.3 Component III: List of data quality attributes

Table 4 presents the template of the *List of Data Quality Attributes* component, which the participants of the focus group validated. It includes eleven fields. Altogether 82 data quality attributes are presented in the concrete implementation of this component. A complete list of all data quality attributes is provided in Appendix B in Tables 13 and 14. A full

[5] https://doi.org/10.7910/DVN/Y6ORUV

[6] https://doi.org/10.7910/DVN/Y6ORUV

[7] *: Found in interview data and literature; **: Found only in interview data; ***: Found only in literature.

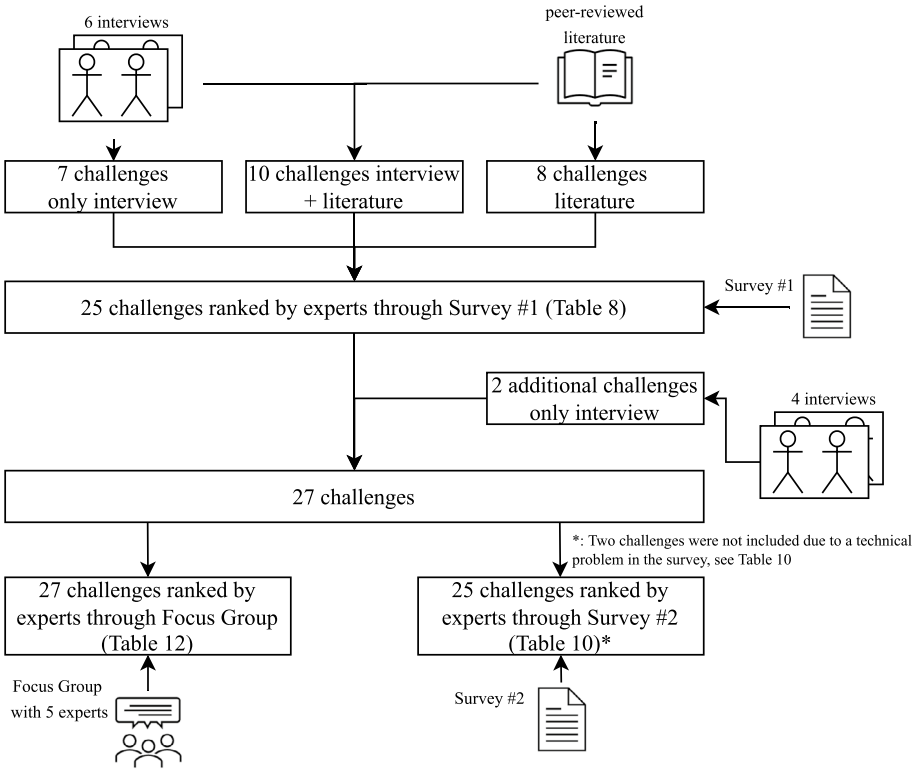

Fig. 3 Challenges identified from various sources and validated using different methods

description of each item is available in the artifact package.[8] Additionally, 30 metrics for different data quality attributes are also presented in the appendix, and a complete description of these metrics is available in the artifact package[1]. E.g., a metric to measure the timeliness of data is the *degree of timeliness*, a ratio between the number of data records received within an acceptable time and the total number of received data records. Furthermore, fields from Planguage, a quality factors notation Gilb (2005), were also adapted for this component.

4.4 Component IV: Solution candidates

Table 5 presents the template of the *Solution Candidates* component. It includes four fields, which were validated by the focus group participants. In the concrete implementation of this component, 13 solution candidates are devised, as depicted in Table 6. Some solution candidates are *Automated Labeling* to solve Low Labeled Data Volume and Manual Data Labeling challenges or *Corroboration of Data with Central Data Repository* to solve the Data Dependent on External Conditions challenge. It should be noted that a single solution can be suitable to solve more than a single challenge.

[8] https://doi.org/10.7910/DVN/Y6ORUV

Table 4 Template for *List of Data Quality Attributes* artifact component (*italics*: Fields from Planguage, not evaluated)

Field	Description	Example
Name	Name of the data quality attribute	Timeliness
Reference	Reference that denotes the identification of the attribute	Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996), Dama International (2017)
Definition	Description of the data quality attribute	Length of time between data availability and the event or phenomenon the data describe (European Commission. Statistical Office of the European Union, 2020) The extent to which the age of the data is appropriate for the task at hand (Wang & Strong, 1996)
Challenges affecting the DQ Attribute	List of challenges that affect the data quality attribute	Data Delay (1, 0.75), Data Drop (0.6, 0.25), Manual Data Collection (0.2, 0.66), Manual Data Labeling (, 0.8)
Metric	Name of the data quality attribute metric	Degree of timeliness
Formula	Formula used to calculate a given metric	Number of data records that are received within an acceptable time / total number of received data records
Scale	The scale of measurement of the metric	The average time (in milliseconds) it takes for data to be received by AI models
Meter	The process or method of measuring the metric	Measured every time a data packet arrives by computer clock
Must	The lowest/highest acceptable threshold value that the metric must have	300 ms
Plan	The value of the metric, if surpassed, is regarded as accepted	100 ms
Wish	The optimum value for the metric	50 ms

Table 5 Template for *solution candidates* artifact component

Field	Description	Example
Name	Name of the solution candidate	Corroboration of data with central data repository
Mitigated challenge	Denotes the challenge(s) a particular solution mitigates	Data dependent on external conditions
Requirement specifications	These are the requirements that should be specified before implementing the solution	1. Define the central data repository, its structure, and address, 2. Define the procedure if the central data repository cannot be contacted, 3. Define the way AI disengagement notification is sent to the user
Implementation details	This presents the step-wise implementation of the solution	The artifact package[a] elaborates the implementation details

[a]https://doi.org/10.7910/DVN/Y6ORUV

Table 6 Concrete solution candidates

Solution candidate	Mitigated challenge
Auto Increasing Sequential Number	Data Drop
Automated Labeling	Low Labeled Data Volume, Manual Data Labeling
Continuous Data Processing	Data Delay
Corroboration of Data with Central Data Repository	Data Dependent on External Conditions
Data Acquisition Solution Task	Data Acquisition
Data Filter	Reliance on Suppliers to Raise Error
Data Level Methods and Algorithm Level Methods	Imbalanced Dataset
Identify Mandatory and Optional Fields	Incomplete Data
Improper Data Transfer Solution Task	Improper Data Transfer
Outlier Techniques	Outlier Data
Pair-wise Attribute Algorithm	Noisy Data
RIASC Tool for Removing Redundancies (RTRR)	Redundant Data
Test Environments	Lack of Variety in Test Environment, Manual Data Collection

In this study, we explored and devised solution candidates. An example of a solution candidate definition is presented in Appendix C. Definitions for all depicted solution candidates can be found in the accompanying artifact package[7].

4.5 In-Depth evaluation of the data quality challenges

Since identifying data quality challenges is one of the goals of this study and a fundamental aspect of the candidate framework, this section presents the results of an in-depth evaluation of the identified data quality challenges.

In order to verify the severity of identified challenges and the validity of the data quality attributes, two surveys and a focus group were conducted. The participants were asked to provide a Likert scale value between 1 and 6 to the challenge sets. They were also asked to rank the individual challenges in the challenge sets. The provided Likert scale values and the rankings were used to compute a *Challenge Score*. The higher the score, the more severe a challenge is compared to other challenges. *Low Labeled Data Volume* challenge (i.e., lack of enough labeled data) had the highest score in both surveys; hence, it is regarded as the most severe challenge.

In addition, during both surveys, the participants were asked to rank the challenge sets on a similar scale. All challenge sets were deemed relevant and showed only minor differences in ranking.

4.5.1 First evaluation survey

Table 7 presents the values of the Likert scale selected for each challenge set by the survey participants. Here, *S1-S6* are the six survey participants. The data is presented in alphabetical order of the challenge set.

Table 7 First survey - Ranking of Challenge Sets

Challenge Set	S1	S2	S3	S4	S5	S6
Data Availability	6	4	5	5	6	3
Data Management	4	3	4	4	6	2
Data Source	5	5	2	3	6	5
Data Structure	2	2	3	1	6	1
Data Trust	3	5	1	2	6	4

Table 8[9] provides the ranking of data quality challenges given by participants of the first survey during the first cycle. Here, *S1-S6* are the six survey participants, $\sum$ depicts the sum of the product of rankings, and *f* is the final normalized *Challenge Score*.

4.5.2 Second evaluation survey

Table 9 presents the values of the Likert scale selected for each challenge set by the second survey participants. Here, *S7-S10* are the four survey participants. The data is presented in alphabetical order of the challenge set.

The ranking of data quality challenges given by participants in the second survey during the third cycle is presented in Table 10[10]. In the table, *S7-S10* are the four survey participants, $\sum$ is the sum of the product of rankings, and *f* is the final normalized *Challenge Score*.

Furthermore, the second survey sent out to the study participants attempted to validate the fields of the templates of the candidate framework components. For all components, every field was evaluated to be appropriate by all survey participants except for two. The field *Sources* in *List of Data Quality Challenges* and the *List of Data Quality Attributes* components are evaluated to be suitable by only 75% and 50% of the survey participants, respectively.

4.5.3 Focus group evaluation

A focus group session was conducted in the third cycle of this study. Five deep learning, data science, and requirement engineering experts participated in the session. Two experts were employed at the case company; three were members of the VEDLIoT research project. Two types of questions were presented during the session. The first type pertains to the ranking of the data quality challenges. The researchers of this thesis study wanted to understand if the experts would rank the challenges differently compared to the ranking of the first cycle of this study. The second question type relates to validating the association between data quality challenges and attributes.

[9] **Note:** In Table 8, *Expensive Procedure* and *Time Consuming* challenges are not included as they were identified only during the second cycle.

[10] **Note:** In Table 10, due to limitation on the number of options provided by the survey tool used (Microsoft Forms), *Manual Data Collection* and *Manual Data Labeling* challenges were combined into a single challenge named *Manual Data Collection and Labeling* for ranking. They are still regarded as separate challenges in the *List of Challenges* artifact component.

Due to a technical error, *Regulatory Compliance* was not included in the second cycle survey. Hence, the calculation of *Challenge Score* ranking disregards it. The disregard is only for calculation of the *Challenge Score*; the challenge is still included in the *List of Challenges* artifact component.

Table 8 Ranking of data quality challenges through the first evaluation survey during the first cycle of the study

Rank	C.Set	Challenge	S1	S2	S3	S4	S5	S6	$\sum$	f
1	D.A.	Low Labeled Data Volume	4	1	4	4	4	4	104	4.333
2	D.So.	Lack of Variety in Test Environment	6	5	6	7	6	7	159	3.786
3	D.A.	Incomplete Data	3	4	2	3	2	3	80	3.333
4	D.So.	Data Dependent on External Conditions	4	7	7	5	7	2	136	3.238
5	D.M.	Manual Data Labeling	10	3	8	10	9	9	193	3.217
6	D.M.	Imbalanced Dataset	7	10	1	9	10	10	178	2.967
7	D.A.	Data Drop	2	3	3	1	3	2	68	2.833
8	D.T.	Incorrect Labeling	5	1	4	5	5	4	80	2.667
9	D.So.	Wrongly-Calibrated / Defective Sensors	3	4	2	6	4	6	111	2.643
10	D.M.	Reliance on Suppliers to Raise Error	6	5	10	5	7	7	155	2.583
11	D.M.	Manual Data Collection	9	2	7	6	8	4	150	2.500
12	D.So.	Noise	5	6	4	3	1	5	103	2.452
13	D.So.	New Data Type	7	3	5	2	5	1	101	2.405
14	D.M.	Large Volume of Data	8	7	5	7	3	8	135	2.250
15	D.M.	Regulatory Compliance	4	8	6	2	6	2	112	1.867
16	D.St.	Unstructured Data	5	5	5	3	2	5	55	1.833
17	D.M.	Data Ownership	5	9	9	3	2	1	109	1.817
18	D.T.	Lack of Good Data from Simulation	3	3	2	3	3	1	54	1.800
19	D.M.	Improper Data Transfer	1	4	4	8	4	6	100	1.667
20	D.A.	Data Delay	1	2	1	2	1	1	38	1.583
21	D.St.	Outlier Data	2	3	1	4	4	1	43	1.433
22	D.M.	Redundant Data	2	6	3	1	5	5	82	1.367
23	D.St.	Incompatible Data Formats	4	4	4	4	1	2	40	1.333
23	D.St.	Data Fragmentation	3	2	2	2	3	4	40	1.333
25	D.M.	Data Acquisition	3	1	2	4	1	3	51	0.850
%	D.M.	Time Consuming	Challenge was found at a later iteration							
%	D.M.	Expensive Procedure	Challenge was found at a later iteration							

Items removed from the final version of the artifact based on the judgment from experts can be found in the artifact package

C.Set Challenge Set, *D.A.* Data Availability, *D.M.* Data Management, *D.So.* Data Source, *D.St.* Data Structure, *D.T.* Data Trust

Unlike in the surveys, the focus group session's ranking portrays the challenges' overall ranking without giving them individual weights and calculating the *Challenge Score*. One of the reasons behind the imposition of a different way is the use of a different tool for the

Table 9 Second survey - Ranking of Challenge Sets

Challenge Set	S7	S8	S9	S10
Data Availability	4	1	6	5
Data Management	4	4	5	2
Data Source	6	3	4	4
Data Structure	3	4	5	2
Data Trust	6	2	6	3

Table 10 Ranking of data quality challenges through the second evaluation survey during the third cycle of the study

Rank	C.Set	Challenge	S7	S8	S9	S10	$\sum$	f
1	D.A.	Low Labeled Data Volume	3	4	4	4	60	3.750
1	D.T.	Incorrect Labeling	3	3	2	3	45	3.750
3	D.So	Wrongly-Calibrated / Defective Sensors	4	2	4	3	58	3.625
3	D.So	Lack of Variety in Test Environment	3	4	3	4	58	3.625
5	D.A.	Incomplete Data	4	3	3	3	52	3.250
6	D.M.	Imbalanced Dataset	8	6	9	10	121	3.025
7	D.So	Noise	2	1	3	1	35	2.917
8	D.M.	Large Volume of Data	4	8	10	4	106	2.650
9	D.St.	Outlier Data	4	3	2	3	40	2.500
10	D.St.	Incompatible Data Formats	3	2	4	1	39	2.438
11	D.M.	Manual Data Collection and Labeling	6	7	7	5	97	2.425
12	D.M.	Data Ownership	2	10	6	9	96	2.400
13	D.M.	Time Consuming	9	9	2	6	94	2.350
14	D.M.	Reliance on Suppliers to Raise Error	7	5	5	8	89	2.225
15	D.A.	Data Drop	2	2	2	2	32	2.000
15	D.St.	Data Fragmentation	1	4	1	4	32	2.000
17	D.M.	Improper Data Transfer	5	3	8	3	78	1.950
18	D.So.	Data Dependent on External Conditions	1	3	2	2	31	1.937
19	D.M.	Expensive Procedure	10	2	3	7	77	1.925
20	D.T.	Lack of Good Data from Simulation	1	2	1	2	22	1.833
21	D.St.	Unstructured Data	2	1	3	2	29	1.813
22	D.So.	New Data Type	2	1	1	1	23	1.438
23	D.M.	Data Acquisition	1	4	4	2	44	1.100
24	D.A.	Data Delay	1	1	1	1	16	1.000
25	D.M.	Redundant Data	3	1	1	1	23	0.575
%	D.M.	Regulatory Compliance	Challenge was not included in ranking, see text.					
%	D.M.	Manual Data Labeling	Challenge was combined with Manual Data Collection, see text.					

Rows with a gray background are the challenges added during the second cycle of the study (Time Consuming, Expensive Procedure)

Regulatory compliance is not included in the following list as there was a technical limitation due to which it was not included in the ranking. Similarly, *manual data collection* and *manual data labeling* were combined into one

C.Set Challenge Set, *D.A.* Data Availability, *D.M.* Data Management, *D.So.* Data Source, *D.St.* Data Structure, *D.T.* Data Trust

focus group. Ranking of the challenge sets using a Likert scale is presented in Table 11. Ranking for challenges in each challenge set is presented in Table 12.

107 data quality challenge-attribute associations were presented for validation during the focus group. The experts regarded only four challenge-attribute associations as not valid (i.e., the initial supposition that the challenges affect the attributes for four of the attributes is not valid in expert opinion). Similarly, for 30 challenge-attribute associations, there was unanimity (i.e., all of the experts in the focus group session regarded a particular challenge as affecting a particular attribute).

Table 11 Ranking of Challenge Sets

Challenge Set	Score
Data Availability	3.5
Data Management	4.0
Data Source	4.0
Data Structure	2.8
Data Trust	5.8

For 45 challenge-attribute associations, more than half, but not all, of the experts in the focus group regarding a particular challenge affect a particular attribute. Similarly, for 26 challenge-attribute associations, more than half, but not all, of the experts in the focus group regarding a particular challenge does not affect a particular attribute. Only for the *Data Delay* challenge, there were two challenge-attribute associations in which half of the experts regarded a particular challenge does affect a particular attribute, and the other half regarded a particular challenge does not affect a particular attribute. This anomaly in data is due to one of the focus group participants not answering the question regarding *Data Delay*.

All data regarding the focus group, including tables outlining the experts' responses, can be found in the data package[11] accompanying this article.

5 Discussion

In this section, we discuss the implications and contributions of our study. We also provide potential threats to the validity of the study.

5.1 Implications and contributions

The study has implications for researchers and practitioners interested in data quality and methods to assess and manage data quality. First, the study provides a candidate framework, which can act as a repository of information regarding data quality. The *Data Quality Workflow* component provides a step-by-step guide of the tasks that could be performed for overall data quality management.

Similarly, the *List of Data Quality Challenges* component provides a tool that can be referred to when designing a system to understand the types of data challenges it could face. *List of Data Quality Attributes* component presents the interested parties with attributes they might want to emphasize more in their systems. For example, a system might prefer data availability more than completeness, or vice versa. The component would help them understand which challenges could affect those attributes. Similarly, interested parties could understand which metrics to focus on and which data to collect to calculate the metric values by using the metrics provided. Mainly, practitioners can record data and compute metrics, which could help them adapt and change their processes if needed.

Likewise, using the *solution candidates* component, they can identify and implement techniques for mitigating the challenges affecting the attributes they prefer most.

[11] https://doi.org/10.7910/DVN/Y6ORUV

Table 12 Ranking of Data Challenges for each Challenge Set based on Expert's knowledge collected in a focus group

Rank	Challenge
Data Availability Challenges	
1	Incomplete Data
2	Low Labeled Data Volume
3	Data Drop
4	Data Delay
Data Management Challenges	
1	Imbalanced Dataset
2	Manual Data Labeling*
3	Regulatory Compliance*
4	Expensive procedure
5	Large Volume of Data
6	Data Acquisition
7	Time Consuming
8	Manual Data Collection*
9	Data Ownership
10	Reliance on Suppliers to Raise Error
11	Improper Data Transfer
12	Redundant Data
Data Source Challenges	
1	Lack of Variety in Test Environment
2	Data Dependent on External Conditions
3	New Data Types
4	Wrongly-calibrated / Defective Sensor
Data Structure Challenges	
1	Unstructured Data
2	Outlier Data
3	Incompatible Data Formats
4	Data Fragmentation
Data Trust Challenges	
1	Incorrect Labeling
2	Lack of Good Data from Simulations
3	Noise

A comparison can be made between the candidate framework proposed in this study and the OpenMDM framework described in Sect. 2. A difference between the two is that OpenMDM provides workflow management of measurement data, whereas CaFDaQAM provides a workflow for overall data quality management. Furthermore, OpenMDM is an Eclipse IDE-based tool, whereas CaFDaQAM could be employed in a programming language-neutral and IDE-neutral fashion.

The candidate framework developed in this study combines various components to present a comprehensive collection of tools to assess and maintain data quality. Those tools include a data quality workflow, templates for identifying and recording data quality challenges and attributes, a list of identified data quality challenges, a list of data quality attributes and

metrics, and a list of solution candidates to many data quality challenges. Prior studies (see Sect. 2) explored a single concept. For example, Cai and Zhu (2015) studied only the aspect of challenges of data quality, Batini et al. (2007) explored steps in data quality risk assessment, and Fletcher (1998) provided an attribute-dimensions matrix. Similarly, Fujii et al. (2020) focused on quality assurance of machine learning-based AI applications and did not touch upon data specifically. Unlike previous studies and frameworks, CaFDaQAM explores the overall data quality management process by explicitly proposing a data quality workflow and providing the necessary tools to apply that workflow. The proposed candidate framework also provides requirements for the individual components.

5.2 Answer to the research questions

In response to the first research question (RQ1), this study identified 27 data quality challenges through interviews and a literature review. We developed a method, *Challenge Score* ranking, to rank and understand the severity of the identified challenges. Furthermore, we verified the identified challenges using surveys and a focus group.

Furthermore, four components were derived, forming the *Candidate Framework for Data Quality Assessment and Maintenance* (CaFDaQAM). A data quality workflow was derived. Similarly, tools were proposed in the form of templates and lists for identifying data quality challenges, identifying, quantifying, and managing data quality attributes, and developing solution candidates for data quality challenges. We validated the candidate framework components using a survey, thus ensuring that they correctly address the need of the stakeholders. Hence, the validated candidate framework components answer the second research question (RQ2).

5.3 Threats to validity

5.3.1 Internal validity

Internal validity is concerned with how different variables affect the result of an experiment. One such threat is researcher bias. We, as researchers, could have introduced biases about the topic of the study. The researchers, for example, could have been biased during collecting data and conducting interviews. In order to mitigate this, two researchers performed thematic coding separately using the same coding technique. They then combined them into a single final set of codes in a joint meeting through discussion.

Similarly, as stated earlier, some challenges were identified through literature review only. However, they were validated by conducting a focus group and surveys. Also, a predefined set of questions was used for the interviews, limiting the discussion during interview sessions. At the end of each interview, the interviewees were asked if any questions that should have been asked were missed. Efforts were made to reduce ambiguity in the questions as much as possible. However, there could still be confusion regarding the questions because of communication gaps.

Likewise, there were a limited number of participants in the interviews, the focus group, and the surveys. Most were from the automated driving sector, which could have skewed the study's result. However, suppose researchers will conduct experiments in the future with the same questionnaire used in this study. In that case, the result could vary if only a few participants are used because those participants might have different experiences and expertise than those consulted during this study.

5.3.2 Reliability

Reliability is associated with the replicability of an experiment or other empirical study, which means future experiments designed in the same fashion as the first experiment should produce the same results as the first experiment. The different versions of interview questions are provided in a replication package so that researchers can track how the research questions evolved based on the participants' responses. The interview questions help researchers to ask similar questions in the future. However, the responses by experts might be different despite being from the same domain and having similar years of experience, which is because they could have different backgrounds and experiences throughout their careers or simply because they can have different perspectives.

5.3.3 Conclusion validity

Conclusion validity deals with the reasonability of the results of an experiment. Because focus group sessions and surveys were conducted to evaluate the artifacts developed in the study, it can be stated that the conclusion of this study is valid. However, the researchers of this study have yet to validate the conclusion with other domain experts, such as healthcare, aerospace, or law enforcement. The artifacts have not been implemented in a real-world context. So, there is scope for future study regarding the real-world implementation of the artifact developed in this thesis.

5.3.4 Generalizability

The study was conducted for a specific sector—automated driving. While our findings and candidate framework can only be generalized beyond this scope with further research, we hope our work can inspire similar concerns in other domains. For instance, quality data is also crucial for critical systems such as healthcare or power grid applications. The candidate framework could be used as a template to identify data quality challenges and mitigate them in such systems. Albeit, modifications in the candidate framework and its components might be warranted for such generalization. Furthermore, we do not claim the generalizability of the identified challenges; we only claim the transferability of the concept that challenges exist in the defined categories.

6 Conclusion

In this study, we have identified data quality challenges that could arise in deep learning systems using an automated driving system case study, thus answering RQ1 of this study. We have identified, analyzed, and evaluated the data quality challenges using interviews and a focus group. The list of challenges acts as one of the components of the candidate framework devised in this study.

The proposed *Candidate Framework for Data Quality Assessment and Maintenance* (CaFDaQAM), its components, and associated templates assist in comprehending data quality challenges, attributes, metrics, and solution candidates. The candidate framework can be used as a tool to improve data quality. It can be used to define data quality requirements for a given system. The proposed templates help create a reference point for identifying data quality issues and defining necessary data attributes. The candidate framework

can help improve the performance of deep learning systems, make better predictions, and reduce the risks that insufficient data quality could pose. Using the information provided by the candidate framework, stakeholders can proactively identify and mitigate the challenges regarding data quality. The candidate framework supports RQ2 of this study.

As future work, researchers can use the candidate framework components as a baseline to further develop a framework. Additional challenges could be identified, or identified challenges could be broken into sub-challenges to explore in detail. In order to make the candidate framework developed in this study generalizable, it can be tested in other fields, such as healthcare. Additional data quality challenges, attributes, and solutions can be identified from different domains.

The candidate framework could also be adopted as an automated tool. Data can be passed through a pipeline in this tool, and different relevant quality aspects of the data can be assessed automatically. Then, quality information can be presented to appropriate stakeholders using various mediums and visualization techniques.

Appendix A. Example of a list of data quality challenges

Here we present an example of a list of data quality challenges. The list was created using the template provided in Table 3. The presented list contains challenges from the category *data availability* challenges. Similar organized lists of challenges for the categories *data management, data sources, data structure*, and *data trust* can be found in the accompanying artifact package.[12]

Data availability challenges

Name: Data Delay

Reference: Interviewee B, Corrales et al. (2016), Kruse et al. (2016)

Description: Data delay can occur during data transmission between different sources and destinations. E.g., a delay can occur in data transmission from sensor to long-term storage, sensor to deep learning functions, and long-term storage to deep learning functions. Similarly, there can also be a delay in receiving a signal sent out by a sensor.

Directly affects AI Functions: 1 “Yes”, 3 “No”

Challenge Score: Survey 1 - 1.583 (Rank 22/31), Survey 2 - 1.000 (Rank 24/25)

Name: Data Drop

Reference: Interviewee D

Description: Some data cycles are dropped now and then, which causes tracking of data to be difficult and disrupts the management and processing of data. Such disruption, in turn, will hinder the training of deep learning models. E.g., dropping three frames in a 30-second clip would mean losing 0.7 s, thus causing a problem for algorithmic correctness.

[12] https://doi.org/10.7910/DVN/Y6ORUV

Directly affects AI Functions: 3 “Yes”, 1 “No”

Challenge Score: Survey 1 - 2.833 (Rank 7/31), Survey 2 - 2.000 (Rank 15/25)

Name: Incomplete Data

Reference: Interviewee E, Corrales et al. (2016),

Description: This challenge is similar to the data drop, as missing data cause both. An incomplete dataset also hinders the training of deep learning models. The difference between data drop and incomplete data is that a record can have all the transmitted bits and yet be incomplete if it does not include some crucial information. However, a data drop occurs when there is a drop in bits.

Directly affects AI functions: 3 “Yes”, 1 “No”

Challenge score: Survey 1 - 3.333 (Rank 3/31), Survey 2 - 3.250 (Rank 5/25)

Name: Low Labeled Data Volume

Reference: Interviewee C

Description: Most of the time, in the training dataset, the volume of the labeled data is significantly lesser than that of the unlabeled data. Since a large volume of data is unlabeled, the unlabeled data is useless, and the deep learning models cannot be adequately trained. E.g., if only 30% of the traffic signs in a scene are labeled, it would be “more difficult for the neural network to learn traffic signs since there are quite a lot of traffic signs among the negative samples.”

Directly affects AI functions: 4 “Yes”, 0 “No”

Challenge score: Survey 1 - 4.333 (Rank 1/31), Survey 2 - 3.750 (Rank 1/25*)

Appendix B: List of data quality attributes

The following Table 13 demonstrates how the template for data quality attributes in Table 4 can be applied to create an organized list of data quality attributes. The following list of data quality attributes and relevant metrics has been compiled and validated with the case company. Table 14 provides the data quality attribute metrics.

Note:

- NA: Not Applicable
- The numbers in the brackets are the weighted average values for the challenge-attribute association calculated from the focus group session and survey 2.
- The first number inside the brackets denotes the weighted average from the focus group results, and the second number denotes the weighted average from survey 2.

Table 13 List of Data Quality Attributes, Their Sources, Definitions, and Association with Data Quality Challenges

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Access Security	Wang and Strong (1996)	*The extent to which access to data can be restricted and hence kept secure.* (Wang & Strong, 1996)	Regulatory Compliance (, 0.66)
Accessibility	Cai and Zhu (2015), Sidi et al. (2012), Wang and Strong (1996), ISO (2008), Statistical Office of the EU (2020), DQ (2017)	*The conditions and modalities by which users can access, use and interpret data.* (Statistical Office of the EU, 2020), *The extent to which data are available or easily and quickly retrievable.* (Wang & Strong, 1996)	Data Acquisition (0.8, 0.66), Data Delay (0.5, 0.5), Data Dependent on External Conditions (1, 1), Data Drop (0.6, 0.5), Data Ownership (, 1), Manual Data Collection (0.2, 0.66)
Accuracy	Interviewees, Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996), ISO (2008)	*The degree to which data values correctly represents real-world entities.* (Data Management Association et al., 2017), *The extent to which data are correct, reliable, and certified free of error.* (Wang & Strong, 1996), *Accuracy of data is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure.* (Statistical Office of the EU, 2020)	Data Dependent on External Conditions (0.6, 0), Data Drop (0.8, 1), Incomplete Data (1, 1), Incorrect Labeling (1, 1), Lack of Good Data from Simulations (0.8, 0.66), Low Labeled Data Volume (0.8, 1), Noise (1, 0.66), Outlier Data (0.4, 0.66), Redundant Data (0.4, 0.33)
Amount of Data	Bobrowski et al. (1998), Wang and Strong (1996)	*The number of facts stored.* (Bobrowski et al., 1998), *The extent to which the quantity or volume of available data is appropriate.* (Wang & Strong, 1996)	NA
Appropriate Amount of Data	Sidi et al. (2012), Wang and Strong (1996)	*The extent to which the quantity or volume of available data is appropriate*. (Wang & Strong, 1996)	NA
Auditability	Cai and Zhu (2015)	*It means that auditors can fairly evaluate data accuracy and integrity within rational time and manpower limits during the data use phase.* (Cai & Zhu, 2015)	Data Ownership (, 0.33)

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Author-ization	Cai and Zhu (2015)	*It refers to whether an individual or organization has the right to use the data.* Cai and Zhu (2015)	NA
Availability	Interviewees, Cai and Zhu (2015), Sidi et al. (2012), ISO (2008)	*The degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use.* (ISO, 2008)	Data Acquisition (1, 0.66), Data Delay (0.25, 0.5), Data Drop (1, 0.75), Incomplete Data (0.6, 0.75), Low Labeled Data Volume (0.2, 0.25)
Believability / Credibility / Reputation	Sidi et al. (2012), Wang and Strong (1996), Cai and Zhu (2015), ISO (2008)	*The degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, commitments).* (ISO, 2008), *The extent to which data are trusted or highly regarded in terms of their source or content.* (Wang & Strong, 1996)	Incomplete Data (1, 1), Incorrect Labeling (1, 1), Lack of Good Data from Simulations (0.6, 1), Outlier Data (0.2, 1), Unstructured Data (, 0)
Clarity / Interpretability / Unambiguous	Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996), Statistical Office of the EU (2020)	*The extent to which data are in an appropriate language and units and the data definitions are clear.* (Wang & Strong, 1996)	Incompatible Data Formats (, 1)
Coherence and Comparability	Statistical Office of the EU (2020)	*Adequacy of statistics to be reliably combined in different ways and for various uses and the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics.* Statistical Office of the EU (2020)	NA
Comment	Statistical Office of the EU (2020)	*Supplementary descriptive text which can be attached to data or metadata.* (Statistical Office of the EU, 2020)	NA

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Complete-ness	Interviewees, Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996)	*Refers to whether all required data is present*. (Data Management Association et al., 2017), *The extent to which data are of sufficient breadth, depth, and scope for the task at hand*. (Wang & Strong, 1996)	Data Delay (0, 0.25), Data Drop (0.8, 1), Improper Data Transfer (0.6, 1), Incomplete Data (1, 1)
Compliance	ISO (2008)	*The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use*. ISO (2008)	Data Ownership (, 1), Regulatory Compliance (, 1)
Conciseness or Concise Representation	Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996)	*The extent to which data are compactly represented without being overwhelming (i.e., brief in presentation, yet complete and to the point)*. (Wang & Strong, 1996)	NA
Confiden-tiality	Statistical Office of the EU (2020), ISO (2008)	*A property of data indicating the extent to which their unauthorised disclosure could be prejudicial or harmful to the interest of the source or other relevant parties*. (Statistical Office of the EU, 2020) *The degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use*. (ISO, 2008)	Data Ownership (, 0.66), Regulatory Compliance (, 0.66)
Consistency / Uniformity	Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), ISO (2008), Data Management Association et al. (2017), DQ (2017)	*Can refer to ensuring that data values are consistently represented within a dataset and between datasets, and consistently associated across datasets*. (Data Management Association et al., 2017), *Measures whether or not data is equivalent across systems or location of storage*. (DQ, 2017)	Data Drop (0.8, 1), Improper Data Transfer (0.6, 0.66), Incompatible Data Formats (, 0.66), Incomplete Data (0.6, 1)

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Consistency and Synchronization	Data Management Association et al. (2017), DQ (2017), Fox et al. (1994)	*A measure of the equivalence of information used in various data stores, applications, and systems, and the processes for making data equivalent.* (Sidi et al., 2012)	see *Consistency / Uniformity* for the challenges affecting this attribute
Consistent Representation / Representational Consistency	Sidi et al. (2012), Wang and Strong (1996)	*The extent to which data are always presented in the same format and are compatible with previous data.* (Wang & Strong, 1996)	Unstructured Data (, 1)
Contact	Statistical Office of the EU (2020)	*Individual or organisational contact points for the data or metadata, including information on how to reach the contact points.* (Statistical Office of the EU, 2020)	Regulatory Compliance (, 0.66)
Correctness	Interviewees, Bobrowski et al. (1998)	*Every set of data stored represents a real world situation.* (Bobrowski et al., 1998)	Data Dependent on External Conditions (0.6, 0.33), Imbalanced Dataset (1, 0.66), Improper Data Transfer (0.6, 0.66), Incomplete Data (0.6,), Incorrect Labeling (1, 1), Low Labeled Data Volume (0.6, 0.75), Noise (0.6, 0.66), Outlier Data (0, 0.33)
Cost and Burden	Statistical Office of the EU (2020)	*Cost associated with the collection and production of a statistical product and burden on respondents*. (Statistical Office of the EU, 2020)	see *Cost Effectiveness* for the challenges affecting this attribute
Cost Effectiveness	Wang and Strong (1996)	*The extent to which the cost of collecting appropriate data is reasonable.* (Wang & Strong, 1996)	Data Acquisition (1, 0.66), Manual Data Collection (1, 0.66), Manual Data Labeling (1,)

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Currency / Currentness	Sidi et al. (2012), Data Management Association et al. (2017), DQ (2017), ISO (2008)	*The measure of whether data values are the most up-to-date version of the information.* (Data Management Association et al., 2017), *The degree to which data has attributes that are of the right age in a specific context of use.* (ISO, 2008)	Data Delay (1, 0.75), Data Drop (0.4, 0.25), Improper Data Transfer (0.4, 1), Incomplete Data (0, 0.75)
Data Coverage	Sidi et al. (2012)	*A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest.* (Sidi et al., 2012)	NA
Data Decay	Sidi et al. (2012)	*A measure of the rate of negative change to data.* (Sidi et al., 2012)	NA
Data Revision	Statistical Office of the EU (2020)	*Any change in a value of a statistic released to the public.* (Statistical Office of the EU, 2020)	NA
Data Specification	Sidi et al. (2012)	*A measure of the existence, completeness, quality and documentation of data standards, data models, business rules, meta data, and reference data.* (Sidi et al., 2012)	NA
Definition / Documentation	Cai and Zhu (2015)	*It consists of data specification, which includes data name, definition, ranges of valid values, standard formats, business rules, etc. Normative data definition improves the degree of data usage.* (Cai & Zhu, 2015)	NA
Duplication	Sidi et al. (2012)	*A measure of unwanted duplication existing within or across systems for a particular field, record, or data set.* (Sidi et al., 2012)	NA

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Ease of Manipulation	Pipino et al. (2002), Sidi et al. (2012)	*The extent to which data is easy to manipulate and apply to different same format.* (Pipino et al., 2002)	NA
Ease of Operation	Wang and Strong (1996)	*The extent to which data are easily managed and manipulated (i.e., updated, moved, aggregated, reproduced, customized).* (Wang & Strong, 1996)	Data Acquisition (0.4, 0.33), Data Ownership (, 0.66), Improper Data Transfer (0.8, 0.66), Manual Data Collection (0.6, 0.66), Manual Data Labeling (0.8,)
Ease of Use and Maintainability	Sidi et al. (2012)	*A measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed.* (McGilvray, 2008), (Sidi et al., 2012)	NA
Effectiveness	Sidi et al. (2012)	*It is the capability of the function to enable users to achieve specified goals with accuracy and completeness in a specified context of use.* (Batini et al., 2009), (Sidi et al., 2012)	NA
Efficiency	Sidi et al. (2012), ISO (2008)	*The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.* (ISO, 2008)	Data Delay (0.75, 0.5), Data Drop (0.6, 0.5), Imbalanced Dataset (0.2, 0), Incomplete Data (0.2, 0.75), Incorrect Labeling (0.2, 0.66), Outlier Data (0.2, 0.33), Unstructured Data (, 0.66)
Fitness	Cai and Zhu (2015)	*It has two-level requirements: 1) the amount of accessed data used by users and 2) the degree to which the data produced matches users' needs in the aspects of indicator definition, elements, classification, etc.* (Cai & Zhu, 2015)	Data Drop (0.4, 0.5), Imbalanced Dataset (1, 0.66), Incomplete Data (0.8, 1), Incorrect Labeling (1, 1), Lack of Good Data From Simulations (0.8, 0.66), Low Labeled Data Volume (0.8, 1), Noise (0.6, 0.66), Outlier Data (0.2, 1)

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Flexibility	Wang and Strong (1996)	*The extent to which data are expandable, adaptable, and easily applied to other needs.* (Wang & Strong, 1996)	Data Drop (0.2, 0.25), Incomplete Data (0.6, 0)
Free of Error	Sidi et al. (2012)	*The extent to which data is correct and reliable* (Pipino et al., 2002), (Sidi et al., 2012)	NA
Frequency of Dissemination	Statistical Office of the EU (2020)	*The time interval at which the statistics are disseminated over a given time period.* (Statistical Office of the EU, 2020)	Regulatory Compliance (, 0.66)
Freshness	Sidi et al. (2012)	*Freshness represents a family of quality factors which each one representing some freshness aspect and having on its metrics.* (Peralta, 2006)	NA
Institutional Mandate	Statistical Office of the EU (2020)	*Law, set of rules or other formal set of instructions assigning responsibility as well as the authority to an organisation for the collection, processing, and dissemination of statistics.* (Statistical Office of the EU, 2020)	Regulatory Compliance (, 1)
Integrity	Cai and Zhu (2015), Sidi et al. (2012), DQ (2017)	*Measures the structural or relational quality of datasets.* (DQ, 2017)	NA
Integrity or Coherence	See *Integrity* and *Coherence*		
Latency	Data Management Association et al. (2017)	*The time between when the data was created and when it was made available for use.* (Data Management Association et al., 2017)	Data Delay (1, 1)
Learnability	Sidi et al. (2012)	*It means the capability of the function to enable to user to learn it.* (Heravizadeh et al., 2009), (Sidi et al., 2012)	NA

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Lineage	DQ (2017)	*Lineage measures whether factual documentation exists about where data came from, how it was transformed, where it went and end-to-end graphical illustration.* (DQ, 2017)	Data Acquisition (1, 1), Data Ownership (, 0.66), Regulatory Compliance (, 0.66)
Metadata	Cai and Zhu (2015)	*With the increase of data sources and data types, because data consumers distort the meaning of common terminology and concepts of data, using data may bring risks. Therefore, data produc- ers need to provide metadata describing different aspects of the datasets to reduce the problems caused by misunderstanding or inconsistencies.* (Cai & Zhu, 2015)	NA
Metadata Update	Statistical Office of the EU (2020)	*The date on which the metadata element was inserted or modified in the database.* (Statistical Office of the EU, 2020)	NA
Navigation	Sidi et al. (2012)	*Extent to which data are easily found and linked to.* (Knight & Burn, 2005), (Sidi et al., 2012)	NA
Objectivity	Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996)	*The extent to which data are unbiased (unprejudiced) and impartial.* (Wang & Strong, 1996)	Data Drop (0.2, 0.75), Incomplete Data (0.2, 1), Incorrect Labeling (0.6, 1), Lack of Good Data from Simulations (0.8, 1), Low Labeled Data Volume (0.6, 0.75), Noise (0.2, 0.66), Outlier Data (0.2, 0.33), Redundant Data (0.2, 0.33)

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Portability	ISO (2008)	*The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another (while) preserving the existing quality in a specific context of use.* (ISO, 2008)	Data Delay (0, 0), Data Drop (0.2, 0.33), Improper Data Transfer (0.8, 0.66), Regulatory Compliance (, 0.66)
Precision	Bobrowski et al. (1998), ISO (2008)	*The degree to which data has attributes that are exact or that provide discrimination in a specific context of use.* (ISO, 2008)	NA
Presentation Quality	Sidi et al. (2012)	*A measure of how information is presented to and collected from does how utilize it. Format and appearance support appropriate use of information.* (McGilvray, 2008), (Sidi et al., 2012)	NA
Punctuality	Statistical Office of the EU (2020)	*Time lag between the actual delivery of the data and the target date when it should have been delivered.* (Statistical Office of the EU, 2020)	NA
Quality Management	Statistical Office of the EU (2020)	*Systems and frameworks in place within an organisation to manage the quality of statistical products and processes.* (Statistical Office of the EU, 2020)	NA
Readability	Cai and Zhu (2015)	*It is defined as the ability of data content to be correctly explained according to known or well-defined terms, attributes, units, codes, abbreviations, or other information.* (Cai & Zhu, 2015)	NA
Reason-ability	Data Management Association et al. (2017)	*Asks whether a data pattern meets expectations.* (Data Management Association et al., 2017)	Data Drop (0.4, 0.5), Incomplete Data (0.8, 0.5)

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Recover-ability	ISO (2008)	*The degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use.* (ISO, 2008)	NA
Reference Period	Statistical Office of the EU (2020)	*The period of time or point in time to which the measured observation is intended to refer.* (Statistical Office of the EU, 2020)	NA
Release Policy	Statistical Office of the EU (2020)	*Rules for disseminating statistical data to all interested parties.* (Statistical Office of the EU, 2020)	Regulatory Compliance (, 0)
Relevance	Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996), Statistical Office of the EU (2020)	*The extent to which data are applicable and helpful for the task at hand.* (Wang & Strong, 1996), *The degree to which statistical information meet current and potential needs of the users.* (Statistical Office of the EU, 2020)	New Data Type (, 0.33)
Reliability	Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012)	*Reliability of the data, defined as the closeness of the initial estimated value to the subsequent estimated value.* (Statistical Office of the EU, 2020)	Data Drop (0.8, 1), Improper Data Transfer (0.8, 0.66), Incomplete Data (0.8, 1), Incorrect Labeling (1, 1)
Represen-tation	DQ (2017)	*Representation measures ease of understanding data, consistency of presentation, appropriate media choice, and availability of documentation (metadata).* (DQ, 2017)	NA

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Safety	Sidi et al. (2012)	*It is the capability of the function to achieve acceptable levels of risk of harm to people, process, property or the environment.* (Heravizadeh et al., 2009), (Sidi et al., 2012)	NA
Security	Sidi et al. (2012)	*Extent to which access to information is restricted appropriately to maintain its security.* (Wang & Strong, 1996), (Sidi et al., 2012)	NA
Statistical Presentation	Statistical Office of the EU (2020)	*Description of the disseminated data which can be displayed to users as tables, graphs or maps.* (Statistical Office of the EU, 2020)	NA
Statistical Processing	Statistical Office of the EU (2020)	This concept and all its sub-concepts are included in ESQRS based (producer) reports. The concept is ESQRS Concept 3. However Sub-concept S.18.5.1 is ESQRS Sub-concept 6.3.4.1 and Sub-concept S.18.6.1 is ESQRS Sub-concept 6.4 (Statistical Office of the EU, 2020).	NA
Structure	Cai and Zhu (2015)	*It refers to the level of difficulty in transforming semi-structured or unstructured data to structured data through technology.* (Cai and Zhu, 2015)	Unstructured Data (, 0.66)
Timeliness	Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012), Wang and Strong (1996), Data Management Association et al. (2017), DQ (2017)	*Length of time between data availability and the event or phenomenon the data describe.* (Statistical Office of the EU, 2020), *The extent to which the age of the data is appropriate for the task at hand.* (Wang & Strong, 1996)	Data Delay (1, 0.75), Data Drop (0.6, 0.25), Manual Data Collection (0.2, 0.66), Manual Data Labeling (, 0.8)

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Timeliness and Availability	Sidi et al. (2012)	*A measure of the degree to which data are current and available for use as specified and in the time frame in which they are expected.* (McGilvray, 2008), (Sidi et al., 2012)	NA
Traceability	Wang and Strong (1996), ISO (2008)	*The extent to which data are well documented, verifiable, and easily attributed to a source.* (Wang & Strong, 1996), *The degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use.* (ISO, 2008)	Data Acquisition (0.8, 1), Data Ownership (, 0.66), Regulatory Compliance (, 0.66)
Transact-ability	Sidi et al. (2012)	*A measure of the degree to which data will produce the desired business transaction or outcome.* (McGilvray, 2008), (Sidi et al., 2012)	NA
Unambi-guous	Bobrowski et al. (1998)	*Each piece of data has a unique meaning.* (Bobrowski et al., 1998)	NA
Understand-ability, ease of understanding	Sidi et al. (2012), Wang and Strong (1996), ISO (2008)	*The degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in (an) appropriate languages, symbols and units in a specific context of use.* (ISO, 2008)	Incomplete Data (0.8, 0.75)
Uniqueness	Data Management Association et al. (2017)	*No entity exists more than once within the dataset.* (Data Management Association et al., 2017)	Redundant Data (, 1)
Unit of Measure	Statistical Office of the EU (2020)	*The unit in which the data values are measured.* (Statistical Office of the EU, 2020)	NA

Table 13 (continued)

DQ Attribute	Source	Definition	Challenge affecting the DQ attribute
Usability	Interviewees, Cai and Zhu (2015), Bobrowski et al. (1998), Sidi et al. (2012)	*Extent to which information is clear and easily used.* (Sidi et al., 2012)	Imbalanced Dataset (1, 1), Incomplete Data (1, 0.5), Incorrect Labeling (1, 1), Low Labeled Data Volume (1, 0.75), Redundant Data (0.6, 0.33), Unstructured Data (, 0.33)
Usefulness	Sidi et al. (2012)	*Extent to which information is applicable and helpful for the task at hand* (Wang & Strong, 1996), (Sidi et al., 2012)	Data Delay (0.5, 0.75), Data Drop (0.4, 0.75), Imbalanced Dataset (1, 1), Incomplete Data (0.8, 0.5), Incorrect Labeling (1, 1), Low Labeled Data Volume (1, 0.75), Lack of Good Data from Simulations (0.8, 0.66), Noise (0.6, 0.66), Redundant Data (0.6,)
Validity	Data Management Association et al. (2017), DQ (2017)	*Refers to whether data values are consistent with a defined domain of values.* (Data Management Association et al., 2017)	Incorrect Labeling (1, 1), Incompatible Data Format (, 0.66), Low Labeled Data Volume (1, 0.75), Unstructured Data (, 0.3)
Value Added	Sidi et al. (2012), Wang and Strong (1996)	*The extent to which data are beneficial and provide advantages from their use.* (Wang & Strong, 1996)	NA
Variety of Data Sources	Wang and Strong (1996)	*The extent to which data are available from several differing data sources.* (Wang & Strong, 1996)	Lack of Good Data from Simulations (0.4, 1)
Volatility	Data Management Association et al. (2017)	*Remain current for a short period.* (Data Management Association et al., 2017)	NA

Table 14 List of Data Quality Attribute Metrics

DQ Attribute	Metric	Formula
Access Security	Degree of security	number of access breaches
Accuracy	Degree of accuracy	number of accurately-labeled data records / total number of data records
Appropriate Amount of Data	Degree of appropriateness	number of minimum amount of data that is required by the system
Availability	Degree of availability	number of successful access to data / total number of access to data
Completeness	Degree of data completeness Degree of mandatory data completeness	number of available data records / total number of data records number of available mandatory data record / total number of mandatory data records
Compliance	Degree of compliance	number of data records that comply with standards / total number of data records
Confidentiality	Degree of security	number of access breaches (by unauthorized users)
Consistency / Uniformity	Degree of consistency	number of consistent data records / total number of data records
Correctness	Degree of correctness	number of correctly-labeled data records / total number of data records
Cost and Burden	Number of over expense	number of instances where the cost exceeded a predefined limit
Cost Effectiveness	Number of over expense	number of instances where the cost exceeded a predefined limit
Currency / Currentness	Degree of currency	number of data records that are latest / total number of data records in a dataset
Data Coverage	Degree of coverage	number of available data / total number of population data
Data Decay	Degree of data decay	number of data records with negative change / total number of data records
Data Revision	Ratio of change in publicly-released information	number of changes in publicly-released information / total number of public releases of information
Data Specification	Ratio of data specification	number of data records that adhere to certain specification / total number of data records
Duplication	Degree of duplication	number of duplicate data records / total number of data records
Effectiveness	Degree of effectiveness	number of goals achieved by AI models / total number of goals of those AI models
Efficiency	Degree of efficiency	number of AI models that perform over the expected level of performance / total number of AI models
Fitness	Degree of fitness	number of data set used by AI models / total number of data set
Frequency of Dissemination	Number of public releases	Number of times information is released in a given time period
Freshness	(see *Currency*)	
Integrity	Degree of integrity	number of uncorrupted data records / total number of data records

Table 14 (continued)

DQ Attribute	Metric	Formula
Latency	Mean latency	sum of latency in given data sets / total number of given data sets
Punctuality	Degree of punctuality	sum of time lag between the actual delivery of data and the target date for given data sets / total number of given data sets
Reasonability	Degree of reasonability	number of data records that meets predefined expectations / total number of data records
Relevance	(see *Fitness*)	
Reliability	Degree of reliability	number of fake data records / total number of data records
Timeliness	Degree of timeliness	number of data records that is received within an acceptable time / total number of received data records
Transactability	(see *Effectiveness*)	
Uniqueness	(see *Duplication*)	
Usefulness	(see *Effectiveness*)	
Variety of Data Sources	Number of data sources	number of data sources
Volatility	Degree of volatility	number of data records that change within a given time period / total number of data records

- If there is no weighted average from either the focus group or survey, the space is left blank. E.g., (, 1) would mean that there is no weighted average from the focus group, but there is a weighted average from survey 2. In the same way, (1,) means vice versa.
- The meaning of weighted average is explained in the main article.

Some data quality attributes do not have an applicable metric. The lack of metrics is that these attributes do not have a tangible numeric value. E.g., *Comment* does not have a numeric value that can be used in devising a metric.

Following is the list of the data quality attributes without a metric.

1. Accessibility
2. Amount of Data
3. Auditability
4. Authorization
5. Believability / Credibility / Reputation
6. Clarity / Interpretability / Unambiguous
7. Coherence and Comparability
8. Comment
9. Conciseness / Concise Representation
10. Consistency and Synchronization
11. Consistent Representation / Representational Consistency
12. Contact
13. Definition / Documentation
14. Ease of Manipulation
15. Ease of Operation
16. Ease of Use and Maintainability
17. Elasticity
18. Flexibility
19. Free of Error
20. Institutional Mandate
21. Learnability
22. Lineage
23. Metadata
24. Metadata Update
25. Navigation
26. Objectivity
27. Portability
28. Precision
29. Presentation Quality
30. Quality Management
31. Readability
32. Recoverability
33. Reference Period
34. Release Policy
35. Representation
36. Resiliency
37. Safety
38. Scalability

39. Security
40. Statistical Presentation
41. Statistical Processing
42. Structure
43. Traceability
44. Unambiguous
45. Understandability / Ease of Understanding
46. Unit of Measure
47. Usability
48. Validity
49. Value Added

Appendix C. Example of a solution candidate

Here we provide an example of a solution candidate. The solution candidate *Continuous Data Processing* has been developed together with the case company. It demonstrates how the template for solution candidates (Table 5) can be applied in practice. Figure 4 shows the flowchart of the solution candidate *Continuous Data Processing*. Altogether 13 solution candidates have been derived in this study. The remaining solution candidates can be found in the supplement material.[13]

Continuous data processing

Mitigated Challenge: Data Delay

Requirement Specifications:

1. Add new fields for departure timestamp and arrival timestamp in the database,
2. Determine an acceptable range of time for data arrival

Implementation Details:

- First, above mentioned requirement specifications, should be completed.
- Then, when the data arrives for processing, check if it is in the initial processing stage.
- CHECK_PIPELINE: If it is, check if there is data in the data pipeline.
  - If there is data in the pipeline, start processing that particular piece of data without waiting for the rest of the data.
  - CHECK_END: If there is no data in the pipeline, check if it is the end of processing.
    * If it is the end of processing, stop.
    * If it is not the end of processing, identify that there is a data delay.
    * Check if the data departure timestamp is there or not.
      · If data departure timestamp exists, compute the total time taken by finding the difference between arrival and departure times.
      · Check if the time taken is within the acceptable range.
      · If it is within the acceptable range, stop.

[13] https://doi.org/10.7910/DVN/Y6ORUV

  - If it is not within the acceptable range, notify appropriate stakeholders about the data delay.

- If it is not the initial stage of processing, check if the stage is mid-processing.
  - If yes, continue from CHECK_PIPELINE.
- If the stage is not mid-processing, continue from CHECK_END.

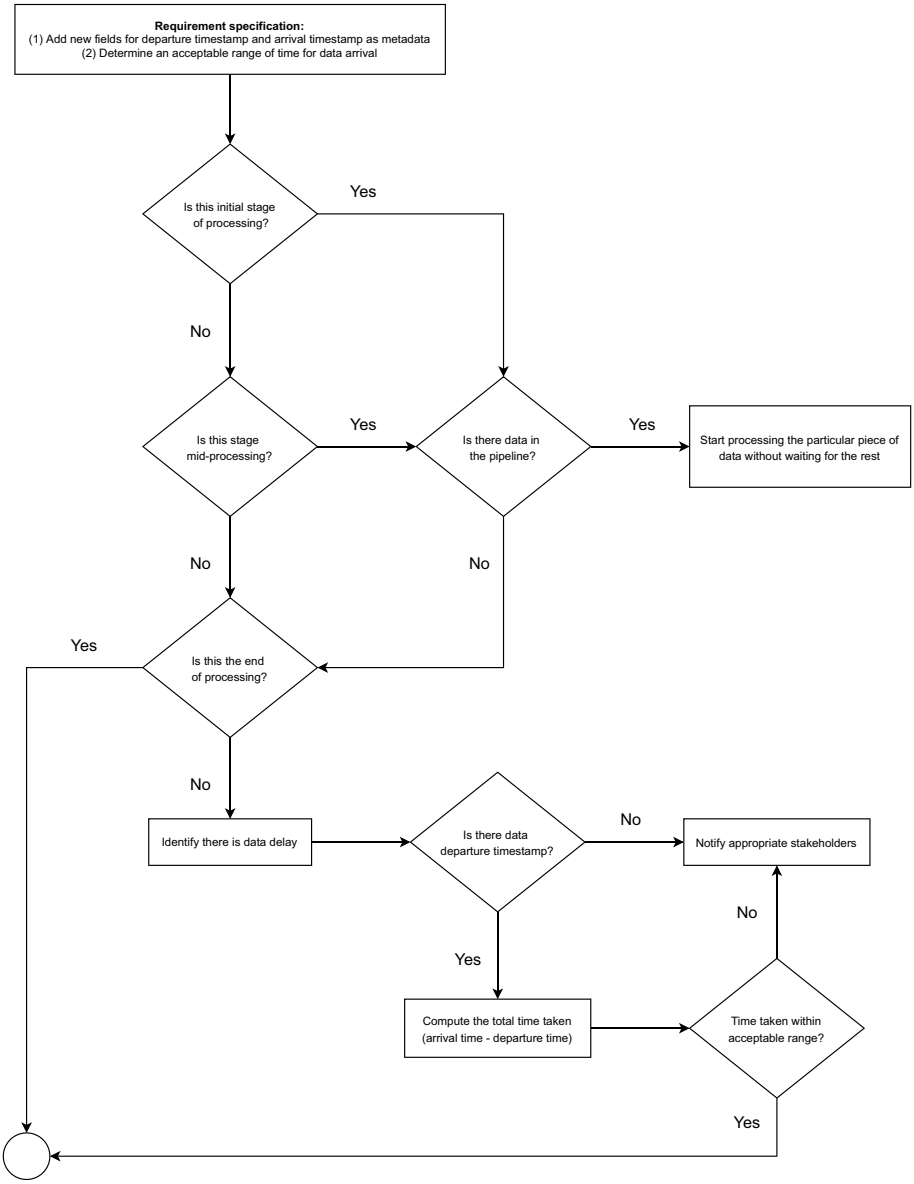

Fig. 4 Flowchart for *Continuous Data Processing* Solution

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 957197.

Author contributions Shameer Kumar Pradhan: Conceptualization, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review and Editing. Hans-Martin Heyn: Conceptualization, Investigation, Methodology, Validation, Resources, Writing - Review and Editing, Supervision. Eric Knauss: Conceptualization, Methodology, Resources, Writing - Review and Editing, Supervision, Project administration, Funding acquisition

Funding Open access funding provided by University of Gothenburg. The research received funding as part of the European Union's Horizon 2020 project "VEDLIoT."

Data availability https://doi.org/10.7910/DVN/Y6ORUV.

Code availability N/A.

Supplementary information The article is accompanied by a replication package which contains a data package and an artifact package. The replication package can be accessed through the Harvard Dataverse at https://doi.org/10.7910/DVN/Y6ORUV. The data package covers the following topics:

- Interview Guide
- Challenge Score
- Focus Group Data
- Survey 2 Data

The artifact package for the article lists and explains the components of the artifact developed in the study. It covers the following components:

- Data Quality Workflow
- Data Quality Challenges
- Data Quality Attributes
- Solution Candidates

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval N/A.

Consent to participate N/A.

Consent for publication N/A.

References

Batini, C., Barone, D., Mastrella, M., Maurino, A., & Ruffini, C. (2007). A framework and a methodology for data quality assessment and monitoring. In *In Proceedings of the 12th International Conference on Information Quality* (pp. 333–346). Cambridge, MA, United States: MIT.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*, 16:1-16:52.

Bobrowski, M., Marré, M., & Yankelevich, D. (1998). A Software Engineering View of Data Quality. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.41.5713 &rep=rep1 &type=pdf.

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*.

Challa, H., Niu, N., & Johnson, R. (2020). Faulty Requirements Made Valuable: On the Role of Data Quality in Deep Learning. In *2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)* (pp. 61–69). Zurich, Switzerland: IEEE.

Corrales, D. C., Ledezma, A. I., & Corrales, J. C. (2016). A systematic review of data quality issues in knowledge discovery tasks. *Revista Ingenierías Universidad de Medellín, 15*, 125–149.

Cox, E. P., III. (1980). The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research, 17*, 407–422.

Dama International. (2017). *Dama-dmbok: Data management body of knowledge* (2nd edition). Denville, NJ, USA: Technics Publications, LLC.

Data Management Association, Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E., & Smith, E. (2017). *DAMA-DMBOK: data management body of knowledge*. Denville, NJ, United States: Technics Publications, LLC.

DQ. (2017). List of Conformed Dimensions of Data Quality | Conformed Dimensions of Data Quality. https://dimensionsofdataquality.com/alldimensions.

European Commission. Statistical Office of the European Union. (2020). *European Statistical System handbook for quality and metadata reports: 2020 edition*. Publications Office, LU. https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf

Farooq, M. B., & de Villiers, C. (2017). Telephonic qualitative research interviews: when to consider them and how to do them. *Meditari Accountancy Research, 25*, 291–316.

Fayyad, J., Jaradat, M. A., Gruyer, D., & Najjaran, H. (2020). Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization: A Review. *Sensors*, *20*, 4220.

Fletcher, F. (1998). A Framework for Addressing Data Quality in Distributed Computing Systems. In *Proceedings of the 1998 International Conference on Information Quality*. MIT.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management, 30*, 9–19.

Fujii, G., Hamada, K., Ishikawa, F., Masuda, S., Matsuya, M., Myojin, T., Nishi, Y., Ogawa, H., Toku, T., Tokumoto, S., Tsuchiya, K., & Ujita, Y. (2020). Guidelines for Quality Assurance of Machine Learning-Based Artificial Intelligence. *International Journal of Software Engineering and Knowledge Engineering, 30*, 1589–1606.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*, 86–92.

Gibbs, G. R. (2007). *Analyzing Qualitative Data*. London, United Kingdom: SAGE Publications, Ltd.

Gilb, T. (2005). *Competitive Engineering: A Handbook For Systems Engineering, Requirements Engineering, and Software Engineering Using Planguage*. Burlington, MA, United States: Elsevier.

Haoues, M., Sellami, A., Ben-Abdallah, H., & Cheikhi, L. (2017). A guideline for software architecture selection based on iso 25010 quality related characteristics. *International Journal of System Assurance Engineering and Management, 8*, 886–909.

Heravizadeh, M., Mendling, J., & Rosemann, M. (2009). Dimensions of Business Processes Quality (QoBP). In D. Ardagna, M. Mecella, & J. Yang (Eds.), *Business Process Management Workshops* (pp. 80–91). Berlin, Heidelberg: Springer.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly, 28*, 75–105.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2020). The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy*, *12*, 1.

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–16).

ISO. (2008). *ISO/IEC 25012:2008*. Technical Report International Organization for Standardization Geneva, Switzerland.

ISO. (2011). *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models*. Geneva: International Organization for Standardization.

Knauss, E. (2021). Constructive Master's Thesis Work in Industry: Guidelines for Applying Design Science Research. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)* (pp. 110–121).

Knight, S.-A., & Burn, J. (2005). Developing a Framework for Assessing Information Quality on the World Wide Web. *Informing Science: The International Journal of an Emerging Transdiscipline, 8*, 159–172.

Kruse, C. S., Goswamy, R., Raval, Y. J., & Marawi, S. (2016). Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Medical Informatics*, *4*.

Madnick, S., Wang, R., & Xian, X. (2014). The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality. *Journal of Management Information Systems, 20*, 41–70.

McGilvray, D. (2008). *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information (TM)*. Academic Press.

McMeekin, N., Wu, O., Germeni, E., & Briggs, A. (2020). How methodological frameworks are being developed: evidence from a scoping review. *BMC Medical Research Methodology*, *20*, 173.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research, 42*, 533–544.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems, 24*, 45–77.

Peralta, V. (2006). *Data Quality Evaluation in Data Integration Systems*. phdthesis Université de Versailles-Saint Quentin en Yvelines ; Université de la République d'Uruguay.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*, 211–218.

Rusk, N. (2016). Deep learning. *Nature Methods*, *13*, 35.

Sandkuhl, K. (2019). Putting AI into Context - Method Support for the Introduction of Artificial Intelligence into Organizations. In *2019 IEEE 21st Conference on Business Informatics (CBI)* (pp. 157–164). volume 01.

Sedgwick, P. (2013). Convenience sampling. *BMJ*, *347*, f6304. Publisher: British Medical Journal Publishing Group Section: Endgames.

Sessions, V., & Valtorta, M. (2006). The effects of data quality on machine learning algorithms. In *11th International Conference on Information Quality* (pp. 485–498). Cambridge, MA, United States: MIT.

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval Knowledge Management* (pp. 300–304).

Skjott Linneberg, M., & Korsgaard, S. (2019). Coding qualitative data: a synthesis guiding the novice. *Qualitative Research Journal, 19*, 259–270.

Statistical Office of the EU. (2020). *European Statistical System handbook for quality and metadata reports: 2020 edition.*. LU: Publications Office.

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision* (pp. 843–852).

Suri, H. (2011). Purposeful Sampling in Qualitative Research Synthesis. *Qualitative Research Journal*, *11*, 63–75. Publisher: Emerald Group Publishing Limited. https://doi.org/10.3316/QRJ1102063

Vogelsang, A., & Borg, M. (2019). Requirements Engineering for Machine Learning: Perspectives from Data Scientists. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)* (pp. 245–251).

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 12*, 5–33.

Ziebiński, A., Cupek, R., Grzechca, D., & Chruszczyk, L. (2017). Review of advanced driver assistance systems (ADAS). *AIP Conference Proceedings*, *1906*.

Shameer Kumar Pradhan is a doctoral candidate in the Business Informatics Research Group at the Faculty of Business Economics of Hasselt University, Belgium. He holds a master's degree in software engineering and technology from Chalmers University of Technology, Sweden. His research focuses on process mining, especially data preprocessing and quality for process mining. His interests also include software engineering, requirements engineering, and digital transformation.

Hans-Martin Heyn is a senior lecturer at the Computer Science and Engineering Department of the University of Gothenburg and Chalmers University of Technology in Sweden. His research concentrates around Software Engineering for AI and distributed cyber-physical systems. He focuses on system architectures, data requirements and runtime monitoring for distributed AI systems. Hans-Martin has more than 5 years of industry experience and previously performed research on distributed sensing for autonomous marine vessels.

Eric Knauss is an Associate Professor (Docent) at the Department of Computer Science and Engineering, Chalmers | University of Gothenburg. His research interests focus on managing requirements and related knowledge in large-scale and distributed software projects to enable digitalization of societies. Research topics include Requirements Engineering, Software Ecosystems, Global and Cross-Organizational Software Development, and Agile Methods. In these areas, Eric has been active for more than 15 years and collaborated with software and systems industry. He has written more than 100 peer-reviewed publication, reviewed for top journals and conferences, and is regularly organizer of scientific events.